PRINCIPLES OF ECONOMETRICS

5th Edition

R. CARTER HILL | WILLIAM E. GRIFFITHS | GUAY C. LIM

WILEY

Principles of Econometrics

Fifth Edition

R. CARTER HILL
Louisiana State University

WILLIAM E. GRIFFITHS
University of Melbourne

GUAY C. LIM
University of Melbourne

WILEY

EDITORIAL DIRECTOR	Michael McDonald
EXECUTIVE EDITOR	Lise Johnson
CHANNEL MARKETING MANAGER	Michele Szczesniak
CONTENT ENABLEMENT SENIOR MANAGER	Leah Michael
CONTENT MANAGEMENT DIRECTOR	Lisa Wojcik
CONTENT MANAGER	Nichole Urban
SENIOR CONTENT SPECIALIST	Nicole Repasky
PRODUCTION EDITOR	Abidha Sulaiman
COVER PHOTO CREDIT	© liuzishan/iStockphoto

This book was set in STIX-Regular 10/12pt by SPi Global and printed and bound by Strategic Content Imaging.

This book is printed on acid free paper. ∞

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

ISBN: 9781118452271 (PBK)
ISBN: 9781119320951 (EVALC)

Library of Congress Cataloging-in-Publication Data

Names: Hill, R. Carter, author. | Griffiths, William E., author. | Lim, G. C. (Guay C.), author.
Title: Principles of econometrics / R. Carter Hill, Louisiana State University, William E. Griffiths, University of Melbourne, Guay C. Lim, University of Melbourn.
Description: Fifth Edition. | Hoboken : Wiley, 2017. | Revised edition of the authors' Principles of econometrics, c2011. | Includes bibliographical references and index. |
Identifiers: LCCN 2017043094 (print) | LCCN 2017056927 (ebook) | ISBN 9781119342854 (pdf) | ISBN 9781119320944 (epub) | ISBN 9781118452271 (paperback) | ISBN 9781119320951 (eval copy)
Subjects: LCSH: Econometrics. | BISAC: BUSINESS & ECONOMICS / Econometrics.
Classification: LCC HB139 (ebook) | LCC HB139 .H548 2018 (print) | DDC 330.01/5195—dc23
LC record available at https://lccn.loc.gov/2017043094

The inside back cover will contain printing identification and country of origin if omitted from this page. In addition, if the ISBN on the back cover differs from the ISBN on this page, the one on the back cover is correct.

10 9 8 7 6 5 4 3 2 1

Carter Hill dedicates this work to his wife, Melissa Waters
Bill Griffiths dedicates this work to Jill, David, and Wendy Griffiths
Guay Lim dedicates this work to Tony Meagher

Brief Contents

Preface

Principles of Econometrics, *Fifth Edition*, is an introductory book for undergraduate students in economics and finance, as well as first-year graduate students in economics, finance, accounting, agricultural economics, marketing, public policy, sociology, law, forestry, and political science. We assume that students have taken courses in the principles of economics and elementary statistics. Matrix algebra is not used, and we introduce and develop calculus concepts in an Appendix. The title *Principles of Econometrics* emphasizes our belief that econometrics should be part of the economics curriculum, in the same way as the principles of microeconomics and the principles of macroeconomics. Those who have been studying and teaching econometrics as long as we have will remember that *Principles of Econometrics* was the title that Henri Theil used for his 1971 classic, which was also published by John Wiley & Sons. Our choice of the same title is not intended to signal that our book is similar in level and content. Theil's work was, and remains, a unique treatise on advanced graduate level econometrics. Our book is an introductory level econometrics text.

Book Objectives

Principles of Econometrics is designed to give students an understanding of why econometrics is necessary and to provide them with a working knowledge of basic econometric tools so that

i. They can apply these tools to modeling, estimation, inference, and forecasting in the context of real-world economic problems.
ii. They can evaluate critically the results and conclusions from others who use basic econometric tools.
iii. They have a foundation and understanding for further study of econometrics.
iv. They have an appreciation of the range of more advanced techniques that exist and that may be covered in later econometric courses.

The book is neither an econometrics cookbook nor is it in a theorem-proof format. It emphasizes motivation, understanding, and implementation. Motivation is achieved by introducing very simple economic models and asking economic questions that the student can answer. Understanding is aided by lucid description of techniques, clear interpretation, and appropriate applications. Learning is reinforced by doing, with clear worked examples in the text and exercises at the end of each chapter.

Overview of Contents

This fifth edition is a major revision in format and content. The chapters contain core material and exercises, while appendices contain more advanced material. Chapter examples are now identified and separated from other content so that they may be easily referenced. From the beginning, we recognize the observational nature of most economic data and modify modeling assumptions accordingly. Chapter 1 introduces econometrics and gives general guidelines for writing an empirical research paper and locating economic data sources. The Probability Primer preceding Chapter 2 summarizes essential properties of random variables and their probability distributions and reviews summation notation. The simple linear regression model is covered in Chapters 2–4, while the multiple regression model is treated in Chapters 5–7. Chapters 8 and 9 introduce econometric problems that are unique to cross-sectional data (heteroskedasticity) and time-series data

(dynamic models), respectively. Chapters 10 and 11 deal with endogenous regressors, the failure of least squares when a regressor is endogenous, and instrumental variables estimation, first in the general case, and then in the simultaneous equations model. In Chapter 12, the analysis of time-series data is extended to discussions of nonstationarity and cointegration. Chapter 13 introduces econometric issues specific to two special time-series models, the vector error correction, and vector autoregressive models, while Chapter 14 considers the analysis of volatility in data and the ARCH model. In Chapters 15 and 16, we introduce microeconometric models for panel data and qualitative and limited dependent variables. In appendices A, B, and C, we introduce math, probability, and statistical inference concepts that are used in the book.

Summary of Changes and New Material

This edition includes a great deal of new material, including new examples and exercises using real data and some significant reorganizations. In this edition, we number examples for easy reference and offer 25–30 new exercises in each chapter. Important new features include

- Chapter 1 includes a discussion of data types and sources of economic data on the Internet. Tips on writing a research paper are given "up front" so that students can form ideas for a paper as the course develops.
- A Probability Primer precedes Chapter 2. This Primer reviews the concepts of random variables and how probabilities are calculated from probability density functions. Mathematical expectation and rules of expected values are summarized for discrete random variables. These rules are applied to develop the concepts of variance and covariance. Calculations of probabilities using the normal distribution are illustrated. New material includes sections on conditional expectation, conditional variance, iterated expectations, and the bivariate normal distribution.
- Chapter 2 now starts with a discussion of causality. We define the population regression function and discuss exogeneity in considerable detail. The properties of the ordinary least squares (OLS) estimator are examined within the framework of the new assumptions. New appendices have been added on the independent variable, covering the various assumptions that might be made about the sampling process, derivations of the properties of the OLS estimator, and Monte Carlo experiments to numerically illustrate estimator properties.
- In Chapter 3, we note that hypothesis test mechanics remain the same under the revised assumptions because test statistics are "pivotal," meaning that their distributions under the null hypothesis do not depend on the data. In appendices, we add an extended discussion of test behavior under the alternative, introduce the noncentral t-distribution, and illustrate test power. We also include new Monte Carlo experiments illustrating test properties when the explanatory variable is random.
- Chapter 4 discusses in detail nonlinear relationships such as the log-log, log-linear, linear-log, and polynomial models. We have expanded the discussion of diagnostic residual plots and added sections on identifying influential observations. The familiar concepts of compound interest are used to motivate several log-linear models. We add an appendix on the concept of mean squared error and the minimum mean squared error predictor.
- Chapter 5 introduces multiple regression in the random-x framework. The Frisch–Waugh–Lovell (FWL) theorem is introduced as a way to help understand interpretation of the multiple regression model and used throughout the remainder of the book. Discussions of the properties of the OLS estimator, and interval estimates and t-tests, are updated. The large sample properties of the OLS estimator, and the delta method, are now introduced within the chapter rather than an appendix. Appendices provide further discussion and Monte Carlo

properties to illustrate the delta method. We provide a new appendix on bootstrapping and its uses.

- Chapter 6 adds a new section on large sample tests. We explain the use of control variables and the difference between causal and predictive models. We revise the discussion of collinearity and include a discussion of influential observations. We introduce nonlinear regression models and nonlinear least squares algorithms are discussed. Appendices are added to discuss the statistical power of F-tests and further uses of the Frisch–Waugh–Lovell theorem.
- Chapter 7 now includes an extensive section on treatment effects and causal modeling in Rubin's potential outcomes framework. We explain and illustrate the interesting regression discontinuity design. An appendix includes a discussion of the important "overlap" assumption.
- Chapter 8 has been reorganized so that the heteroskedasticity robust variance of the OLS estimator appears before testing. We add a section on how model specification can ameliorate heteroskedasticity in some applications. We add appendices to explain the properties of the OLS residuals and another to explain alternative robust sandwich variance estimators. We present Monte Carlo experiments to illustrate the differences.
- Chapter 9 has been reorganized and streamlined. The initial section introduces the different ways that dynamic elements can be added to the regression model. These include using finite lag models, infinite lag models, and autoregressive errors. We carefully discuss autocorrelations, including testing for autocorrelation and representing autocorrelations using a correlogram. After introducing the concepts of stationarity and weak dependence, we discuss the general notions of forecasting and forecast intervals in the context of autoregressive distributed lag (ARDL) models. Following these introductory concepts, there are details of estimating and using alternative models, covering such topics as choosing lag lengths, testing for Granger causality, the Lagrange multiplier test for serial correlation, and using models for policy analysis. We provide very specific sets of assumptions for time-series regression models and outline how heteroskedastic and autocorrelation consistent, robust, standard errors are used. We discuss generalized least squares estimation of a time-series regression model and its relation to nonlinear least squares regression. A detailed discussion of the infinite lag model and how to use multiplier analysis is provided. An appendix contains details of the Durbin–Watson test.
- Chapter 10 on endogeneity problems has been streamlined because the concept of random explanatory variables is now introduced much earlier in the book. We provide further analysis of weak instruments and how weak instruments adversely affect the precision of IV estimation. The details of the Hausman test are now included in the chapter.
- Chapter 11 now adds Klein's Model I as an example.
- Chapter 12 includes more details of deterministic trends and unit roots. The section on unit root testing has been restructured so that each Dickey–Fuller test is more fully explained and illustrated with an example. Numerical examples of ARDL models with nonstationary variables that are, and are not, cointegrated have been added.
- The data in Chapter 13 have been updated and new exercises added.
- Chapter 14 mentions further extensions of ARCH volatility models.
- Chapter 15 has been restructured to give priority to how panel data can be used to cope with the endogeneity caused by unobserved heterogeneity. We introduce the advantages of having panel data using the first difference estimator, and then discuss the within/fixed effects estimator. We provide an extended discussion of cluster robust standard errors in both the OLS and fixed effects model. We discuss the Mundlak version of the Hausman test for endogeneity. We give brief mention to how to extend the use of panel data in several ways.

- The Chapter 16 discussion of binary choice models is reorganized and expanded. It now includes brief discussions of advanced topics such as binary choice models with endogenous explanatory variables and binary choice models with panel data. We add new appendices on random utility models and latent variable models.
- Appendix A includes new sections on second derivatives and finding maxima and minima of univariate and bivariate functions.
- Appendix B includes new material on conditional expectations and conditional variances, including several useful decompositions. We include new sections on truncated random variables, including the truncated normal and Poisson distributions. To facilitate discussions of test power, we have new sections on the noncentral t-distribution, the noncentral Chi-square distribution, and the noncentral F-distribution. We have included an expanded new section on the log-normal distribution.
- Appendix C content does not change a great deal, but 20 new exercises are included.
- Statistical Tables for the Standard Normal cumulative distribution function, the t-distribution and Chi-square distribution critical values for selected percentiles, the F-distribution critical values for the 95th and 99th percentiles, and the Standard Normal density function values appear in Appendix D.
- A useful "cheat sheet" of essential formulas is provided at the authors' website, www.principlesofeconometrics.com, rather than inside the covers as in the previous edition.

For Instructors: Suggested Course Plans

Principles of Econometrics, Fifth Edition is suitable for one or two semester courses at the undergraduate or first year graduate level. Some suitable plans for alternative courses are as follows:

- One-semester survey course: Sections P.1–P.6.2 and P.7; Sections 2.1–2.9; Chapters 3 and 4; Sections 5.1–5.6; Sections 6.1–6.5; Sections 7.1–7.3; Sections 8.1–8.4 and 8.6; Sections 9.1–9.4.2 and 9.5–9.5.1.
- One-semester survey course enhancements for Master's or Ph.D.: Include Appendices for Chapters 2–9.
- Two-semester survey second course, cross-section emphasis: Section P.6; Section 2.10; Section 5.7; Section 6.6; Sections 7.4–7.6; Sections 8.5 and 8.6.3–8.6.5; Sections 10.1–10.4; Sections 15.1–15.4; Sections 16.1–16.2 and 16.6;
- Two-semester survey second course, time series emphasis: Section P.6; Section 2.10; Section 5.7; Section 6.6; Sections 7.4–7.6; Sections 8.5 and 8.6.3–8.6.5; Section 9.5; Sections 10.1–10.4; Sections 12.1–12.5; Sections 13.1–13.5; Sections 14.1–14.4;
- Two-semester survey course enhancements for Master's or Ph.D.: Include Appendices from Chapters 10, Chapter 11, Appendices 15A–15B, Sections 16.3–16.5 and 16.7, Appendices 16A–16D, Book Appendices B and C.

Computer Supplement Books

There are several computer supplements to *Principles of Econometrics, Fifth Edition*. The supplements are not versions of the text and cannot substitute for the text. They use the examples in the text as a vehicle for learning the software. We show how to use the software to get the answers for each example in the text.

- *Using EViews for the Principles of Econometrics, Fifth Edition,* by William E. Griffiths, R. Carter Hill, and Guay C. Lim [ISBN 9781118469842]. This supplementary book presents the EViews 10 [www.eviews.com] software commands required for the examples in *Principles of Econometrics* in a clear and concise way. It includes many illustrations that are student friendly. It is useful not only for students and instructors who will be using this software as part of their econometrics course but also for those who wish to learn how to use EViews.
- *Using Stata for the Principles of Econometrics, Fifth Edition*, by Lee C. Adkins and R. Carter Hill [ISBN 9781118469873]. This supplementary book presents the Stata 15.0 [www.stata.com] software commands required for the examples in *Principles of Econometrics*. It is useful not only for students and instructors who will be using this software as part of their econometrics course but also for those who wish to learn how to use Stata. Screen shots illustrate the use of Stata's drop-down menus. Stata commands are explained and the use of "do-files" illustrated.
- *Using SAS for the Principles of Econometrics, Fifth Edition*, by Randall C. Campbell and R. Carter Hill [ISBN 9781118469880]. This supplementary book gives SAS 9.4 [www.sas.com] software commands for econometric tasks, following the general outline of *Principles of Econometrics, Fifth Edition*. It includes enough background material on econometrics so that instructors using any textbook can easily use this book as a supplement. The volume spans several levels of econometrics. It is suitable for undergraduate students who will use "canned" SAS statistical procedures, and for graduate students who will use advanced procedures as well as direct programming in SAS's matrix language; the latter is discussed in chapter appendices.
- *Using Excel for Principles of Econometrics, Fifth Edition,* by Genevieve Briand and R. Carter Hill [ISBN 9781118469835]. This supplement explains how to use Excel to reproduce most of the examples in *Principles of Econometrics*. Detailed instructions and screen shots are provided explaining both the computations and clarifying the operations of Excel. Templates are developed for common tasks.
- *Using GRETL for Principles of Econometrics, Fifth Edition*, by Lee C. Adkins. This free supplement, readable using Adobe Acrobat, explains how to use the freely available statistical software GRETL (download from http://gretl.sourceforge.net). Professor Adkins explains in detail, and using screen shots, how to use GRETL to replicate the examples in *Principles of Econometrics.* The manual is freely available at www.learneconometrics.com/gretl.html.
- *Using R for Principles of Econometrics, Fifth Edition*, by Constantin Colonescu and R. Carter Hill. This free supplement, readable using Adobe Acrobat, explains how to use the freely available statistical software *R* (download from https://www.r-project.org/). The supplement explains in detail, and using screen shots, how to use *R* to replicate the examples in *Principles of Econometrics, Fifth Edition.* The manual is freely available at https://bookdown.org/ccolonescu/RPOE5/.

Data Files

Data files for the book are provided in a variety of formats at the book website www.wiley.com/college/hill. These include

- ASCII format (*.dat). These are text files containing only data.
- Definition files (*.def). These are text files describing the data file contents, with a listing of variable names, variable definitions, and summary statistics.
- EViews (*.wf1) workfiles for each data file.
- Excel (*.xls) workbooks for each data file, including variable names in the first row.

- Comma separated values (*.csv) files that can be read into almost all software.
- Stata (*.dta) data files.
- SAS (*.sas7bdat) data files.
- GRETL (*.gdt) data files.
- R (*.rdata) data files.

The author website www.principlesofeconometrics.com includes a complete list of the data files and where they are used in the book.

Additional Resources

The book website www.principlesofeconometrics.com includes

- Individual data files in each format as well as ZIP files containing data in compressed format.
- Book errata.
- Brief answers to odd number problems. These answers are also provided on the book website at www.wiley.com/college/hill.
- Additional examples with solutions. Some extra examples come with complete solutions so that students will know what a good answer looks like.
- Tips on writing research papers.

Resources for Instructors

For instructors, also available at the website www.wiley.com/college/hill are

- Complete solutions, in both Microsoft Word and *.pdf formats, to *all* exercises in the text.
- PowerPoint slides and PowerPoint Viewer.

Acknowledgments

Authors Hill and Griffiths want to acknowledge the gifts given to them over the past 40 years by mentor, friend, and colleague George Judge. Neither this book nor any of the other books we have shared in the writing of would have ever seen the light of day without his vision and inspiration.

We also wish to give thanks to the many students and faculty who have commented on the fourth edition of the text and contributed to the fifth edition. This list includes Alejandra Breve Ferrari, Alex James, Alyssa Wans, August Saibeni, Barry Rafferty, Bill Rising, Bob Martin, Brad Lewis, Bronson Fong, David Harris, David Iseral, Deborah Williams, Deokrye Baek, Diana Whistler, Emma Powers, Ercan Saridogan, Erdogan Cevher, Erika Haguette, Ethan Luedecke, Gareth Thomas, Gawon Yoon, Genevieve Briand, German Altgelt, Glenn Sueyoshi, Henry McCool, James Railton, Jana Ruimerman, Jeffery Parker, Joe Goss, John Jackson, Julie Leiby, Katharina Hauck, Katherine Ramirez, Kelley Pace, Lee Adkins, Matias Cattaneo, Max O'Krepki, Meagan McCollum, Micah West, Michelle Savolainen, Oystein Myrland, Patrick Scholten, Randy Campbell, Regina Riphahn, Sandamali Kankanamge, Sergio Pastorello, Shahrokh Towfighi, Tom Fomby, Tong Zeng, Victoria Pryor, Yann Nicolas, and Yuanbo Zhang. In the book errata we acknowledge those who have pointed out our errors.

R. Carter Hill
William E. Griffiths
Guay C. Lim

Table of Contents

List of Examples

CHAPTER 1

An Introduction to Econometrics

1.1 Why Study Econometrics?

Econometrics is fundamental for economic measurement. However, its importance extends far beyond the discipline of economics. Econometrics is a set of research tools also employed in the business disciplines of accounting, finance, marketing, and management. It is used by social scientists, specifically researchers in history, political science, and sociology. Econometrics plays an important role in such diverse fields as forestry and agricultural economics. This breadth of interest in econometrics arises in part because economics is the foundation of business analysis and is the core social science. Thus, research methods employed by economists, which includes the field of econometrics, are useful to a broad spectrum of individuals.

Econometrics plays a special role in the training of economists. As a student of economics, you are learning to "think like an economist." You are learning economic concepts such as opportunity cost, scarcity, and comparative advantage. You are working with economic models of supply and demand, macroeconomic behavior, and international trade. Through this training you become a person who better understands the world in which we live; you become someone who understands how markets work, and the way in which government policies affect the marketplace.

If economics is your major or minor field of study, a wide range of opportunities is open to you upon graduation. If you wish to enter the business world, your employer will want to know the answer to the question, "What can you do for me?" Students taking a traditional economics curriculum answer, "I can think like an economist." While we may view such a response to be powerful, it is not very specific and may not be very satisfying to an employer who does not understand economics.

The problem is that a gap exists between what you have learned as an economics student and what economists actually do. Very few economists make their livings by studying economic theory alone, and those who do are usually employed by universities. Most economists, whether they work in the business world or for the government, or teach in universities, engage in economic analysis that is in part "empirical." By this we mean that they use economic data to estimate economic relationships, test economic hypotheses, and predict economic outcomes.

Studying econometrics fills the gap between being "a student of economics" and being "a practicing economist." With the econometric skills you will learn from this book, including how to work with econometric software, you will be able to elaborate on your answer to the employer's question above by saying "I can predict the sales of your product." "I can estimate the effect on your sales if your competition lowers its price by $1 per unit." "I can test whether your new ad campaign is actually increasing your sales." These answers are music to an employer's ears, because they reflect your ability to think like an economist and to analyze economic data.

Such pieces of information are keys to good business decisions. Being able to provide your employer with useful information will make you a valuable employee and increase your odds of getting a desirable job.

On the other hand, if you plan to continue your education by enrolling in graduate school or law school, you will find that this introduction to econometrics is invaluable. If your goal is to earn a master's or Ph.D. degree in economics, finance, data analytics, data science, accounting, marketing, agricultural economics, sociology, political science, or forestry, you will encounter more econometrics in your future. The graduate courses tend to be quite technical and mathematical, and the forest often gets lost in studying the trees. By taking this introduction to econometrics you will gain an overview of what econometrics is about and develop some "intuition" about how things work before entering a technically oriented course.

1.2 What Is Econometrics About?

At this point we need to describe the nature of econometrics. It all begins with a theory from your field of study—whether it is accounting, sociology, or economics—about how important variables are related to one another. In economics we express our ideas about relationships between economic variables using the mathematical concept of a function. For example, to express a relationship between income and consumption, we may write

$$CONSUMPTION = f(INCOME)$$

which says that the level of consumption is *some* function, $f(\bullet)$, of income.

The demand for an individual commodity—say, the Honda Accord—might be expressed as

$$Q^d = f(P, P^s, P^c, INC)$$

which says that the quantity of Honda Accords demanded, Q^d, is a function $f(P, P^s, P^c, INC)$ of the price of Honda Accords P, the price of cars that are substitutes P^s, the price of items that are complements P^c (like gasoline), and the level of income INC.

The supply of an agricultural commodity such as beef might be written as

$$Q^s = f\left(P, P^c, P^f\right)$$

where Q^s is the quantity supplied, P is the price of beef, P^c is the price of competitive products in production (e.g., the price of hogs), and P^f is the price of factors or inputs (e.g., the price of corn) used in the production process.

Each of the above equations is a general economic model that describes how we visualize the way in which economic variables are interrelated. Economic models of this type *guide our economic analysis*.

Econometrics allows us to go further than knowing that certain economic variables are interrelated, or even the direction of a relationship. Econometrics allows us to assign magnitudes to questions about the interrelationships between variables. One aspect of econometrics is **prediction** or **forecasting**. If we know the value of *INCOME*, what will be the magnitude of *CONSUMPTION*? If we have values for the prices of Honda Accords, their substitutes and complements, and income, how many Honda Accords will be sold? Similarly, we could ask how much beef would be supplied given values of the variables on which its supply depends.

A second contribution of econometrics is to enable us to say **how much** a change in one variable affects another. If the price for Honda Accords is increased, by how much will quantity demanded decline? If the price of beef goes up, by how much will quantity supplied increase? Finally, econometrics contributes to our understanding of the interrelationships between variables by giving us the ability to **test** the validity of hypothesized relationships.

Econometrics is about how we can use theory and data from economics, business, and the social sciences, along with tools from statistics, to predict outcomes, answer "how much" type questions, and test hypotheses.

1.2.1 Some Examples

Consider the problem faced by decision makers in a central bank. In the United States, the Federal Reserve System and, in particular, the Chair of the Board of Governors of the FRB must make decisions about interest rates. When prices are observed to rise, suggesting an increase in the inflation rate, the FRB must make a decision about whether to dampen the rate of growth of the economy. It can do so by raising the interest rate it charges its member banks when they borrow money (the discount rate) or the rate on overnight loans between banks (the federal funds rate). Increasing these rates sends a ripple effect through the economy, causing increases in other interest rates, such as those faced by would-be investors, who may be firms seeking funds for capital expansion or individuals who wish to buy consumer durables like automobiles and refrigerators. This has the economic effect of increasing costs, and consumers react by reducing the quantity of the durable goods demanded. Overall, aggregate demand falls, which slows the rate of inflation. These relationships are suggested by economic theory.

The real question facing the Chair is "*How much* should we increase the discount rate to slow inflation and yet maintain a stable and growing economy?" The answer will depend on the responsiveness of firms and individuals to increases in the interest rates and to the effects of reduced investment on gross national product (GNP). The key elasticities and multipliers are called **parameters**. The values of economic parameters are unknown and must be estimated using a sample of economic data when formulating economic policies.

Econometrics is about how to best estimate economic parameters given the data we have. "Good" econometrics is important since errors in the estimates used by policymakers such as the FRB may lead to interest rate corrections that are too large or too small, which has consequences for all of us.

Every day, decision-makers face "how much" questions similar to those facing the FRB Chair:

- A city council ponders the question of how much violent crime will be reduced if an additional million dollars is spent putting uniformed police on the street.
- The owner of a local Pizza Hut must decide how much advertising space to purchase in the local newspaper and thus must estimate the relationship between advertising and sales.
- Louisiana State University must estimate how much enrollment will fall if tuition is raised by $300 per semester and thus whether its revenue from tuition will rise or fall.
- The CEO of Proctor & Gamble must predict how much demand there will be in 10 years for the detergent Tide and how much to invest in new plant and equipment.
- A real estate developer must predict by how much population and income will increase to the south of Baton Rouge, Louisiana, over the next few years and whether it will be profitable to begin construction of a gambling casino and golf course.
- You must decide how much of your savings will go into a stock fund and how much into the money market. This requires you to make predictions of the level of economic activity, the rate of inflation, and interest rates over your planning horizon.
- A public transportation council in Melbourne, Australia, must decide how an increase in fares for public transportation (trams, trains, and buses) will affect the number of travelers who switch to car or bike and the effect of this switch on revenue going to public transportation.

To answer these questions of "how much," decision-makers rely on information provided by empirical economic research. In such research, an economist uses economic theory and reasoning to construct relationships between the variables in question. Data on these variables are collected and econometric methods are used to estimate the key underlying parameters and to make predictions. The decision-makers in the above examples obtain their "estimates" and "predictions" in different ways. The FRB has a large staff of economists to carry out econometric analyses. The CEO of Proctor & Gamble may hire econometric consultants to provide the firm with projections of sales. You may get advice about investing from a stock broker, who in turn is provided with econometric projections made by economists working for the parent company. Whatever the source of your information about "how much" questions, it is a good bet that there is an economist involved who is using econometric methods to analyze data that yield the answers.

In the next section, we show how to introduce parameters into an economic model and how to convert an economic model into an econometric model.

1.3 The Econometric Model

What is an econometric model, and where does it come from? We will give you a general overview, and we may use terms that are unfamiliar to you. Be assured that before you are too far into this book, all the terminology will be clearly defined. In an econometric model we must first realize that economic relations are not exact. Economic theory does not claim to be able to predict the specific behavior of any individual or firm, but rather describes the *average* or *systematic* behavior of *many* individuals or firms. When studying car sales we recognize that the *actual* number of Hondas sold is the sum of this systematic part and a random and unpredictable component e that we will call a **random error**. Thus, an **econometric model** representing the sales of Honda Accords is

$$Q^d = f(P, P^s, P^c, INC) + e$$

The random error e accounts for the many factors that affect sales that we have omitted from this simple model, and it also reflects the intrinsic uncertainty in economic activity.

To complete the specification of the econometric model, we must also say something about the form of the algebraic relationship among our economic variables. For example, in your first economics courses quantity demanded was depicted as a *linear* function of price. We extend that assumption to the other variables as well, making the systematic part of the demand relation

$$f(P, P^s, P^c, INC) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC$$

The corresponding econometric model is

$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC + e$$

The coefficients β_1, β_2,…,β_5 are unknown **parameters** of the model that we estimate using economic data and an econometric technique. The functional form represents a hypothesis about the relationship between the variables. In any particular problem, one challenge is to determine a functional form that is compatible with economic theory and the data.

In every econometric model, whether it is a demand equation, a supply equation, or a production function, there is a systematic portion and an unobservable random component. The systematic portion is the part we obtain from economic theory, and includes an assumption about the functional form. The random component represents a "noise" component, which obscures our understanding of the relationship among variables, and which we represent using the random variable e.

We use the econometric model as a basis for **statistical inference**. Using the econometric model and a sample of data, we make inferences concerning the real world, learning something in the process. The ways in which statistical inference are carried out include the following:

- **Estimating** economic parameters, such as elasticities, using econometric methods

- **Predicting** economic outcomes, such as the enrollment in two-year colleges in the United States for the next 10 years
- **Testing** economic hypotheses, such as the question of whether newspaper advertising is better than store displays for increasing sales

Econometrics includes all of these aspects of statistical inference. As we proceed through this book, you will learn how to properly estimate, predict, and test, given the characteristics of the data at hand.

1.3.1 Causality and Prediction

A question that often arises when specifying an econometric model is whether a relationship can be viewed as both causal and predictive or only predictive. To appreciate the difference, consider an equation where a student's grade in Econometrics *GRADE* is related to the proportion of class lectures that are skipped *SKIP*.

$$GRADE = \beta_1 + \beta_2 SKIP + e$$

We would expect β_2 to be negative: the greater the proportion of lectures that are skipped, the lower the grade. But, can we say that skipping lectures **causes** grades to be lower? If lectures are captured by video, they could be viewed at another time. Perhaps a student is skipping lectures because he or she has a demanding job, and the demanding job does not leave enough time for study, and this is the underlying cause of a poor grade. Or, it might be that skipping lectures comes from a general lack of commitment or motivation, and this is the cause of a poor grade. Under these circumstances, what can we say about the equation that relates *GRADE* to *SKIP*? We can still call it a predictive equation. *GRADE* and *SKIP* are (negatively) correlated and so information about *SKIP* can be used to help predict *GRADE*. However, we cannot call it a causal relationship. Skipping lectures does not cause a low grade. The parameter β_2 does not convey the direct causal effect of skipping lectures on grade. It also includes the effect of other variables that are omitted from the equation and correlated with *SKIP*, such as hours spent studying or student motivation.

Economists are frequently interested in parameters that can be interpreted as causal. Honda would like to know the direct effect of a price change on the sales of their Accords. When there is technological improvement in the beef industry, the price elasticities of demand and supply have important implications for changes in consumer and producer welfare. One of our tasks will be to see what assumptions are necessary for an econometric model to be interpreted as causal and to assess whether those assumptions hold.

An area where predictive relationships are important is in the use of "big data." Advances in computer technology have led to storage of massive amounts of information. Travel sites on the Internet keep track of destinations you have been looking at. Google targets you with advertisements based on sites that you have been surfing. Through their loyalty cards, supermarkets keep data on your purchases and identify sale items relevant for you. Data analysts use big data to identify predictive relationships that help predict our behavior.

In general, the type of data we have impacts on the specification of an econometric model and the assumptions that we make about it. We turn now to a discussion of different types of data and where they can be found.

1.4 How Are Data Generated?

In order to carry out statistical inference we must have data. Where do data come from? What type of real processes generate data? Economists and other social scientists work in a complex world in which data on variables are "observed" and rarely obtained from a controlled experiment. This makes the task of learning about economic parameters all the more difficult. Procedures for using such data to answer questions of economic importance are the subject matter of this book.

1.4.1 Experimental Data

One way to acquire information about the unknown parameters of economic relationships is to conduct or observe the outcome of an experiment. In the physical sciences and agriculture, it is easy to imagine controlled experiments. Scientists specify the values of key control variables and then observe the outcome. We might plant similar plots of land with a particular variety of wheat, and then vary the amounts of fertilizer and pesticide applied to each plot, observing at the end of the growing season the bushels of wheat produced on each plot. Repeating the experiment on N plots of land creates a sample of N observations. Such controlled experiments are rare in business and the social sciences. A key aspect of experimental data is that the values of the explanatory variables can be fixed at specific values in repeated trials of the experiment.

One business example comes from marketing research. Suppose we are interested in the weekly sales of a particular item at a supermarket. As an item is sold it is passed over a scanning unit to record the price and the amount that will appear on your grocery bill. But at the same time, a data record is created, and at every point in time the price of the item and the prices of all its competitors are known, as well as current store displays and coupon usage. The prices and shopping environment are controlled by store management, so this "experiment" can be repeated a number of days or weeks using the same values of the "control" variables.

There are some examples of planned experiments in the social sciences, but they are rare because of the difficulties in organizing and funding them. A notable example of a planned experiment is Tennessee's Project Star.[1] This experiment followed a single cohort of elementary school children from kindergarten through the third grade, beginning in 1985 and ending in 1989. In the experiment children and teachers were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. The objective was to determine the effect of small classes on student learning, as measured by student scores on achievement tests. We will analyze the data in Chapter 7 and show that small classes significantly increase performance. This finding will influence public policy toward education for years to come.

1.4.2 Quasi-Experimental Data

It is useful to distinguish between "pure" experimental data and "quasi"-experimental data. A pure experiment is characterized by random assignment. In the example where varying amounts of fertilizer and pesticides are applied to plots of land for growing wheat, the different applications of fertilizer and pesticides are randomly assigned to different plots. In Tennessee's Project Star, students and teachers are randomly assigned to different sized classes with and without a teacher's aide. In general, if we have a control group and a treatment group, and we want to examine the effect of a policy intervention or treatment, pure experimental data are such that individuals are randomly assigned to the control and treatment groups.

Random assignment is not always possible however, particularly when dealing with human subjects. With quasi-experimental data, allocation to the control and treatment groups is not random but based on another criterion. An example is a study by Card and Krueger that is studied in more detail in Chapter 7. They examined the effect of an increase in New Jersey's minimum wage in 1992 on the number of people employed in fast-food restaurants. The treatment group was fast-food restaurants in New Jersey. The control group was fast-food restaurants in eastern Pennsylvania where there was no change in the minimum wage. Another example is the effect on spending habits of a change in the income tax rate for individuals above a threshold income. The treatment group is the group with incomes above the threshold. The control group is those with incomes below the threshold. When dealing with quasi-experimental data, one must be aware that the effect of the treatment may be confounded with the effect of the criterion for assignment.

[1] See https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766 for program description, public use data, and extensive literature.

1.4.3 Nonexperimental Data

An example of nonexperimental data is survey data. The Public Policy Research Lab at Louisiana State University (www.survey.lsu.edu) conducts telephone and mail surveys for clients. In a telephone survey, numbers are selected randomly and called. Responses to questions are recorded and analyzed. In such an environment, data on all variables are collected simultaneously, and the values are neither fixed nor repeatable. These are nonexperimental data.

Such surveys are carried out on a massive scale by national governments. For example, the Current Population Survey (CPS)[2] is a monthly survey of about 50,000 households conducted by the U.S. Bureau of the Census. The survey has been conducted for more than 50 years. The CPS website says "CPS data are used by government policymakers and legislators as important indicators of our nation's economic situation and for planning and evaluating many government programs. They are also used by the press, students, academics, and the general public." In Section 1.8 we describe some similar data sources.

1.5 Economic Data Types

Economic data comes in a variety of "flavors." In this section we describe and give an example of each. In each example, be aware of the different data characteristics, such as the following:

1. Data may be collected at various levels of aggregation:
   - *micro*—data collected on individual economic decision-making units such as individuals, households, and firms.
   - *macro*—data resulting from a pooling or aggregating over individuals, households, or firms at the local, state, or national levels.
2. Data may also represent a flow or a stock:
   - *flow*—outcome measures over a period of time, such as the consumption of gasoline during the last quarter of 2018.
   - *stock*—outcome measured at a particular point in time, such as the quantity of crude oil held by ExxonMobil in its U.S. storage tanks on November 1, 2018, or the asset value of the Wells Fargo Bank on July 1, 2018.
3. Data may be quantitative or qualitative:
   - *quantitative*—outcomes such as prices or income that may be expressed as numbers or some transformation of them, such as real prices or per capita income.
   - *qualitative*—outcomes that are of an "either-or" situation. For example, a consumer either did or did not make a purchase of a particular good, or a person either is or is not married.

1.5.1 Time-Series Data

A **time-series** is data collected over discrete intervals of time. Examples include the annual price of wheat in the United States and the daily price of General Electric stock shares. Macroeconomic data are usually reported in monthly, quarterly, or annual terms. Financial data, such as stock prices, can be recorded daily, or at even higher frequencies. The key feature of time-series data is that the same economic quantity is recorded at a regular time interval.

For example, the annual real gross domestic product (GDP) for the United States is depicted in Figure 1.1. A few values are given in Table 1.1. For each year, we have the recorded value. The data are annual, or yearly, and have been "deflated" by the Bureau of Economic Analysis to billions of real 2009 dollars.

[2] www.census.gov/cps/

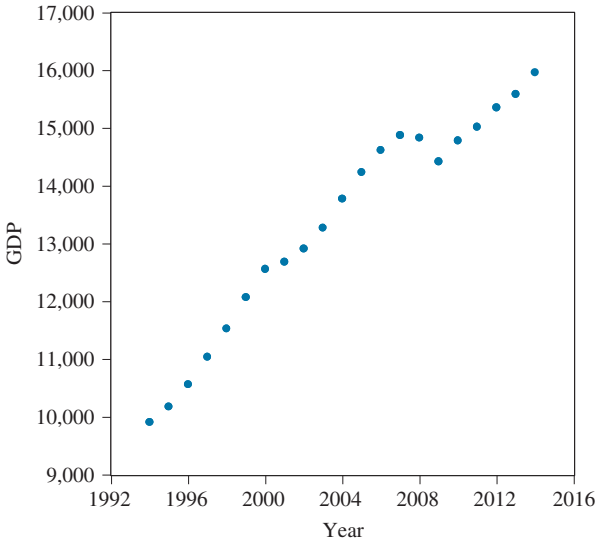


FIGURE 1.1 **Real U.S. GDP, 1994–2014.**[3]

TABLE 1.1 **U.S. Annual GDP (Billions of Real 2009 Dollars)**

Year	GDP
2006	14,613.8
2007	14,873.7
2008	14,830.4
2009	14,418.7
2010	14,783.8
2011	15,020.6
2012	15,354.6
2013	15,583.3
2014	15,961.7

1.5.2 Cross-Section Data

A cross-section of data is collected across sample units in a particular time period. Examples are income by counties in California during 2016 or high school graduation rates by state in 2015. The "sample units" are individual entities and may be firms, persons, households, states, or countries. For example, the CPS reports results of personal interviews on a monthly basis, covering items such as employment, unemployment, earnings, educational attainment, and income. In Table 1.2, we report a few observations from the March 2013 survey on the variables *RACE, EDUCATION, SEX*, and *WAGE* (hourly wage rate).[4] There are many detailed questions asked of the respondents.

[3] Source: www.bea.gov/national/index.htm

[4] In the actual raw data, the variable descriptions are coded differently to the names in Table 1.2. We have used shortened versions for convenience.

TABLE 1.2 **Cross-Section Data: CPS, March 2013**

	Variables			
Individual	***RACE***	***EDUCATION***	***SEX***	***WAGE***
1	White	Assoc Degree	Male	10.00
2	White	Master's Degree	Male	60.83
3	Black	Bachelor's Degree	Male	17.80
4	White	High School Graduate	Female	30.38
5	White	Master's Degree	Male	12.50
6	White	Master's Degree	Female	49.50
7	White	Master's Degree	Female	23.08
8	Black	Assoc Degree	Female	28.95
9	White	Some College, No Degree	Female	9.20

1.5.3 Panel or Longitudinal Data

A "panel" of data, also known as "longitudinal" data, has observations on individual micro-units that are followed over time. For example, the Panel Study of Income Dynamics (PSID)[5] describes itself as "a nationally representative longitudinal study of nearly 9000 U.S. families. Following the same families and individuals since 1969, the PSID collects data on economic, health, and social behavior." Other national panels exist, and many are described at "Resources for Economists," at www.rfe.org.

To illustrate, data from two rice farms[6] are given in Table 1.3. The data are annual observations on rice farms (or firms) over the period 1990–1997.

The key aspect of panel data is that we observe each micro-unit, here a farm, for a number of time periods. Here we have amount of rice produced, area planted, labor input, and fertilizer use. If we have the same number of time period observations for each micro-unit, which is the case here, we have a **balanced panel**. Usually the number of time-series observations is small relative to the number of micro-units, but not always. The Penn World Table[7] provides purchasing power parity and national income accounts converted to international prices for 182 countries for some or all of the years 1950–2014.

1.6 The Research Process

Econometrics is ultimately a research tool. Students of econometrics plan to do research or they plan to read and evaluate the research of others, or both. This section provides a frame of reference and guide for future work. In particular, we show you the role of econometrics in research.

Research is a process, and like many such activities, it flows according to an orderly pattern. Research is an adventure, and can be *fun!* Searching for an answer to your question, seeking new knowledge, is very addictive—for the more you seek, the more new questions you will find.

A research project is an opportunity to investigate a topic that is important to you. Choosing a good research topic is essential if you are to complete a project successfully. A starting point is the question "What are my interests?" Interest in a particular topic will add pleasure to the

[5]http://psidonline.isr.umich.edu

[6]These data were used by O'Donnell, C.J. and W.E. Griffiths (2006), Estimating State-Contingent Production Frontiers, *American Journal of Agricultural Economics*, 88(1), 249–266.

[7]www.rug.nl/ggdc/productivity/pwt

TABLE 1.3 Panel Data from Two Rice Farms

FARM	YEAR	PROD	AREA	LABOR	FERT
1	1990	7.87	2.50	160	207.5
1	1991	7.18	2.50	138	295.5
1	1992	8.92	2.50	140	362.5
1	1993	7.31	2.50	127	338.0
1	1994	7.54	2.50	145	337.5
1	1995	4.51	2.50	123	207.2
1	1996	4.37	2.25	123	345.0
1	1997	7.27	2.15	87	222.8
2	1990	10.35	3.80	184	303.5
2	1991	10.21	3.80	151	206.0
2	1992	13.29	3.80	185	374.5
2	1993	18.58	3.80	262	421.0
2	1994	17.07	3.80	174	595.7
2	1995	16.61	4.25	244	234.8
2	1996	12.28	4.25	159	479.0
2	1997	14.20	3.75	133	170.0

research effort. Also, if you begin working on a topic, other questions will usually occur to you. These new questions may put another light on the original topic or may represent new paths to follow that are even more interesting to you. The idea may come after lengthy study of all that has been written on a particular topic. You will find that "inspiration is 99% perspiration." That means that after you dig at a topic long enough, a new and interesting question will occur to you. Alternatively, you may be led by your natural curiosity to an interesting question. Professor Hal Varian[8] suggests that you look for ideas outside academic journals—in newspapers, magazines, etc. He relates a story about a research project that developed from his shopping for a new TV set.

By the time you have completed several semesters of economics classes, you will find yourself enjoying some areas more than others. For each of us, specialized areas such as health economics, economic development, industrial organization, public finance, resource economics, monetary economics, environmental economics, and international trade hold a different appeal. If you find an area or topic in which you are interested, consult the *Journal of Economic Literature* (*JEL*) for a list of related journal articles. The *JEL* has a classification scheme that makes isolating particular areas of study an easy task. Alternatively, type a few descriptive words into your favorite search engine and see what pops up.

Once you have focused on a particular idea, begin the research process, which generally follows steps like these:

1. Economic theory gives us a way of thinking about the problem. Which economic variables are involved, and what is the possible direction of the relationship(s)? Every research project, given the initial question, begins by building an economic model and listing the questions (hypotheses) of interest. More questions will arise during the research project, but it is good to list those that motivate you at the project's beginning.
2. The working economic model leads to an econometric model. We must choose a functional form and make some assumptions about the nature of the error term.

[8] Varian, H. How to Build an Economic Model in Your Spare Time, *The American Economist*, 41(2), Fall 1997, pp. 3–10.

3. Sample data are obtained and a desirable method of statistical analysis chosen, based on initial assumptions and an understanding of how the data were collected.
4. Estimates of the unknown parameters are obtained with the help of a statistical software package, predictions are made, and hypothesis tests are performed.
5. Model diagnostics are performed to check the validity of assumptions. For example, were all of the right-hand side explanatory variables relevant? Was an adequate functional form used?
6. The economic consequences and the implications of the empirical results are analyzed and evaluated. What economic resource allocation and distribution results are implied, and what are their policy-choice implications? What remaining questions might be answered with further study or with new and better data?

These steps provide some direction for what must be done. However, research always includes some surprises that may send you back to an earlier point in your research plan or that may even cause you to revise it completely. Research requires a sense of urgency, which keeps the project moving forward, the patience not to rush beyond careful analysis, and the willingness to explore new ideas.

1.7 Writing an Empirical Research Paper

Research rewards you with new knowledge, but it is incomplete until a research paper or report is written. The process of writing forces the distillation of ideas. In no other way will your depth of understanding be so clearly revealed. When you have difficulty explaining a concept or thought, it may mean that your understanding is incomplete. Thus, writing is an integral part of research. We provide this section as a building block for future writing assignments. Consult it as needed. You will find other tips on writing economics papers on the book website, www.principlesofeconometrics.com.

1.7.1 Writing a Research Proposal

After you have selected a specific topic, it is a good idea to write up a brief project summary, or proposal. Writing it will help to focus your thoughts about what you really want to do. Show it to your colleagues or instructor for preliminary comments. The summary should be short, usually no longer than 500 words, and should include the following:

1. A concise statement of the problem
2. Comments on the information that is available, with one or two key references
3. A description of the research design that includes
   a. the economic model
   b. the econometric estimation and inference methods
   c. data sources
   d. estimation, hypothesis testing, and prediction procedures, including the econometric software and version used
4. The potential contribution of the research

1.7.2 A Format for Writing a Research Report

Economic research reports have a standard format in which the various steps of the research project are discussed and the results interpreted. The following outline is typical.

1. *Statement of the Problem* The place to start your report is with a summary of the questions you wish to investigate as well as why they are important and who should be interested in the results. This introductory section should be nontechnical and should motivate the reader to continue reading the paper. It is also useful to map out the contents of the following sections of the report. This is the first section to work on and also the last. In today's busy world, the reader's attention must be garnered very quickly. A clear, concise, well-written introduction is a must and is arguably the most important part of the paper.
2. *Review of the Literature* Briefly summarize the relevant literature in the research area you have chosen and clarify how your work extends our knowledge. By all means, cite the works of others who have motivated your research, but keep it brief. You do not have to survey everything that has been written on the topic.
3. *The Economic Model* Specify the economic model that you used and define the economic variables. State the model's assumptions and identify hypotheses that you wish to test. Economic models can get complicated. Your task is to explain the model clearly, but as briefly and simply as possible. Don't use unnecessary technical jargon. Use simple terms instead of complicated ones when possible. Your objective is to display the quality of your thinking, not the extent of your vocabulary.
4. *The Econometric Model* Discuss the econometric model that corresponds to the economic model. Make sure you include a discussion of the variables in the model, the functional form, the error assumptions, and any other assumptions that you make. Use notation that is as simple as possible, and do not clutter the body of the paper with long proofs or derivations; these can go into a technical appendix.
5. *The Data* Describe the data you used, as well as the source of the data and any reservations you have about their appropriateness.
6. *The Estimation and Inference Procedures* Describe the estimation methods you used and why they were chosen. Explain hypothesis testing procedures and their usage. Indicate the software used and the version, such as Stata 15 or EViews 10.
7. *The Empirical Results and Conclusions* Report the parameter estimates, their interpretation, and the values of test statistics. Comment on their statistical significance, their relation to previous estimates, and their economic implications.
8. *Possible Extensions and Limitations of the Study* Your research will raise questions about the economic model, data, and estimation techniques. What future research is suggested by your findings, and how might you go about performing it?
9. *Acknowledgments* It is appropriate to recognize those who have commented on and contributed to your research. This may include your instructor, a librarian who helped you find data, or a fellow student who read and commented on your paper.
10. *References* An alphabetical list of the literature you cite in your study, as well as references to the data sources you used.

Once you've written the first draft, use your computer's spell-check software to check for spelling errors. Have a friend read the paper, make suggestions for clarifying the prose, and check your logic and conclusions. Before you submit the paper, you should eliminate as many errors as possible. Your work should look good. Use a word processor, and be consistent with font sizes, section headings, style of footnotes, references, and so on. Often software developers provide templates for term papers and theses. A little searching for a good paper layout before beginning is a good idea. Typos, missing references, and incorrect formulas can spell doom for an otherwise excellent paper. Some do's and don'ts are summarized nicely, and with good humor, by Deidre N. McClosky in *Economical Writing*, 2nd edition (Prospect Heights, IL: Waveland Press, Inc., 1999).

While it is not a pleasant topic to discuss, you should be aware of the rules of **plagiarism**. You must not use someone else's words as if they were your own. If you are unclear about what you can and cannot use, check with the style manuals listed in the next paragraph, or consult

your instructor. Your university may provide a plagiarism-checking software, such as Turnitin or iThenticate, that will compare your paper to millions of online sources and look for problem areas. There are some free online versions as well. The paper should have clearly defined sections and subsections. The pages, equations, tables, and figures should be numbered. References and footnotes should be formatted in an acceptable fashion. A style guide is a good investment. Two classics are the following:

- *The Chicago Manual of Style*, 16th edition, is available online and in other formats.
- *A Manual for Writers of Research Papers, Theses, and Dissertations: Chicago Style for Students and Researchers*, 8th edition, by Kate L. Turabian; revised by Wayne C. Booth, Gregory G. Colomb, and Joseph M Williams (2013, University of Chicago Press).

1.8 Sources of Economic Data

Economic data are much easier to obtain since the development of the World Wide Web. In this section we direct you to some places on the Internet where economic data are accessible. During your study of econometrics, browse some of the sources listed to gain some familiarity with data availability.

1.8.1 Links to Economic Data on the Internet

There are a number of fantastic sites on the World Wide Web for obtaining economic data.

Resources for Economists (RFE) www.rfe.org is a primary gateway to resources on the Internet for economists. This excellent site is the work of Bill Goffe. Here you will find links to sites for economic data and sites of general interest to economists. The **Data** link has these broad data categories:

- *U.S. Macro and Regional Data* Here you will find links to various data sources such as the Bureau of Economic Analysis, Bureau of Labor Statistics, *Economic Reports of the President*, and the Federal Reserve Banks.
- *Other U.S. Data* Here you will find links to the U.S. Census Bureau, as well as links to many panel and survey data sources. The gateway to U.S. government agencies is FedStats (fedstats.sites.usa.gov). Once there, click on *Agencies* to see a complete list of U.S. government agencies and links to their homepages.
- *World and Non-U.S. Data* Here there are links to world data, such as at the CIA World Factbook and the Penn World Tables, as well as international organizations such as the Asian Development Bank, the International Monetary Fund, the World Bank, and so on. There are also links to sites with data on specific countries and sectors of the world.
- *Finance and Financial Markets* Here are links to sources of U.S. and world financial data on variables such as exchange rates, interest rates, and share prices.
- *Journal Data and Program Archives* Some economic journals post data used in articles. Links to these journals are provided here. (Many of the articles in these journals will be beyond the scope of undergraduate economics majors.)

National Bureau of Economic Research (NBER) www.nber.org/data provides access to a great amount of data. There are headings for

- Macro Data
- Industry Productivity and Digitalization Data
- International Trade Data

- Individual Data
- Healthcare Data—Hospitals, Providers, Drugs, and Devices
- Demographic and Vital Statistics
- Patent and Scientific Papers Data
- Other Data

Economagic Some websites make extracting data relatively easy. For example, Economagic (www.economagic.com) is an excellent and easy-to-use source of macro time series (some 100,000 series available). The data series are easily viewed in a copy and paste format, or graphed.

1.8.2 Interpreting Economic Data

In many cases it is easier to obtain economic data than it is to understand the meaning of the data. It is essential when using macroeconomic or financial data that you understand the definitions of the variables. Just what is the index of leading economic indicators? What is included in personal consumption expenditures? You may find the answers to some questions like these in your textbooks. Another resource you might find useful is *A Guide to Everyday Economic Statistics*, 7th edition, by Gary E. Clayton and Martin Gerhard Giesbrecht, (Boston: Irwin/McGraw-Hill, 2009). This slender volume examines how economic statistics are constructed, and how they can be used.

1.8.3 Obtaining the Data

Finding a data source is not the same as obtaining the data. Although there are a great many easy-to-use websites, "easy-to-use" is a relative term. The data will come packaged in a variety of formats. It is also true that there are many, many variables at each of these websites. A primary challenge is identifying the specific variables that you want, and what exactly they measure. The following examples are illustrative.

The Federal Reserve Bank of St. Louis [9] has a system called **FRED** (Federal Reserve Economic Data). Under "Categories" there are links to financial variables, population and labor variables, national accounts, and many others. Data on these variables can be downloaded in a number of formats. For reading the data, you may need specific knowledge of your statistical software. Accompanying *Principles of Econometrics, 5e*, are computer manuals for Excel, EViews, Stata, SAS, R, and Gretl to aid this process. See the publisher website www.wiley.com/college/hill, or the book website at www.principlesofeconometrics.com for a description of these aids.

The CPS (www.census.gov/cps) has a tool called **DataFerrett**. This tool will help you find and download data series that are of particular interest to you. There are tutorials that guide you through the process. Variable descriptions, as well as the specific survey questions, are provided to aid in your selection. It is somewhat like an Internet shopping site. Desired series are "ticked" and added to a "Shopping Basket." Once you have filled your basket, you download the data to use with specific software. Other Web-based data sources operate in this same manner. One example is the PSID.[10] The Penn World Tables[11] offer data downloads in both Excel and Stata formats.

You can expect to find massive amounts of readily available data at the various sites we have mentioned, but there is a learning curve. You should not expect to find, download, and process the data without considerable work effort. Being skilled with Excel and statistical software is a must if you plan to regularly use these data sources.

[9] https://fred.stlouisfed.org

[10] http://psidonline.isr.umich.edu

[11] www.rug.nl/ggdc/productivity/pwt

Probability Primer

LEARNING OBJECTIVES

Remark

Learning Objectives and *Keywords* sections will appear at the beginning of each chapter. We urge you to think about, and possibly write out answers to the questions, and make sure you recognize and can define the keywords. If you are unsure about the questions or answers, consult your instructor. When examples are requested in *Learning Objectives* sections, you should think of examples *not* in the book.

Based on the material in this primer, you should be able to

1. Explain the difference between a random variable and its values, and give an example.
2. Explain the difference between discrete and continuous random variables, and give examples of each.
3. State the characteristics of a probability density function (*pdf*) for a discrete random variable, and give an example.
4. Compute probabilities of events, given a discrete probability function.
5. Explain the meaning of the following statement: ''The probability that the discrete random variable takes the value 2 is 0.3.''
6. Explain how the *pdf* of a continuous random variable is different from the *pdf* of a discrete random variable.
7. Show, geometrically, how to compute probabilities given a *pdf* for a continuous random variable.
8. Explain, intuitively, the concept of the mean, or expected value, of a random variable.
9. Use the definition of expected value for a discrete random variable to compute expectations, given a *pdf* $f(x)$ and a function $g(X)$ of X.
10. Define the variance of a discrete random variable, and explain in what sense the values of a random variable are more spread out if the variance is larger.
11. Use a joint *pdf* (table) for two discrete random variables to compute probabilities of joint events and to find the (marginal) *pdf* of each individual random variable.
12. Find the conditional *pdf* for one discrete random variable given the value of another and their joint *pdf*.
13. Work with single and double summation notation.
14. Give an intuitive explanation of statistical independence of two random variables, and state the conditions that must hold to prove statistical independence. Give examples of two independent random variables and two dependent random variables.

15. Define the covariance and correlation between two random variables, and compute these values given a joint probability function of two discrete random variables.
16. Find the mean and variance of a sum of random variables.
17. Use Statistical Table 1, Cumulative Probabilities for the Standard Normal Distribution, and your computer software to compute probabilities involving normal random variables.
18. Use the Law of Iterated Expectations to find the expected value of a random variable.

KEYWORDS

conditional expectation
conditional *pdf*
conditional probability
continuous random variable
correlation
covariance
cumulative distribution function
discrete random variable
expected value
experiment
indicator variable
iterated expectation
joint probability density function
marginal distribution
mean
normal distribution
population
probability
probability density function
random variable
standard deviation
standard normal distribution
statistical independence
summation operations
variance

We assume that you have had a basic probability and statistics course. In this primer, we review some essential probability concepts. Section P.1 defines discrete and continuous random variables. Probability distributions are discussed in Section P.2. Section P.3 introduces joint probability distributions, defines conditional probability and **statistical independence**. In Section P.4, we digress and discuss operations with summations. In Section P.5, we review the properties of probability distributions, paying particular attention to expected values and variances. In Section P.6, we discuss the important concept of **conditioning**, and how knowing the value of one variable might provide information about, or help us predict, another variable. Section P.7 summarizes important facts about the normal probability distribution. In Appendix B, "Probability Concepts," are enhancements and additions to this material.

P.1 Random Variables

Benjamin Franklin is credited with the saying "The only things certain in life are death and taxes." While not the original intent, this bit of wisdom points out that almost everything we encounter in life is uncertain. We do not know how many games our football team will win next season. You do not know what score you will make on the next exam. We don't know what the stock market index will be tomorrow. These events, or outcomes, are uncertain, or random. **Probability** gives us a way to talk about possible outcomes.

A **random variable** is a variable whose value is unknown until it is observed; in other words, it is a variable that is not perfectly predictable. Each random variable has a set of possible values it can take. If W is the number of games our football team wins next year, then W can take the values 0, 1, 2, …, 13, if there are a maximum of 13 games. This is a **discrete random variable** since it can take only a limited, or **countable**, number of values. Other examples of discrete random variables are the number of computers owned by a randomly selected household, and the number of times you will visit your physician next year. A special case occurs when a random variable can only be one of two possible values—for example, in a phone survey, if you are asked if you are a college graduate or not, your answer can only be "yes" or "no." Outcomes like this can be characterized by an **indicator variable** taking the values one if yes or zero if no. Indicator variables are discrete and are used to represent qualitative characteristics such as sex (male or female) or race (white or nonwhite).

The U.S. GDP is yet another example of a random variable, because its value is unknown until it is observed. In the third quarter of 2014 it was calculated to be 16,164.1 billion dollars. What the value will be in the second quarter of 2025 is unknown, and it cannot be predicted perfectly. GDP is measured in dollars and it *can* be counted in whole dollars, but the value is so large that counting individual dollars serves no purpose. For practical purposes, GDP can take any value in the interval zero to infinity, and it is treated as a **continuous random variable**. Other common macroeconomic variables, such as interest rates, investment, and consumption, are also treated as continuous random variables. In finance, stock market indices, like the Dow Jones Industrial Index, are also treated as continuous. The key attribute of these variables that makes them continuous is that they can take any value in an interval.

P.2 Probability Distributions

Probability is usually defined in terms of **experiments**. Let us illustrate this in the context of a simple experiment. Consider the objects in Table P.1 to be a population of interest. In statistics and econometrics, the term **population** is an important one. A population is a group of objects, such as people, farms, or business firms, having something in common. The population is a complete set and is the focus of an analysis. In this case the population is the set of ten objects shown in Table P.1. Using this population, we will discuss some probability concepts. In an **empirical analysis,** a sample of observations is collected from the population of interest, and using the sample observations we make inferences about the population.

If we were to select one cell from the table at random (imagine cutting the table into 10 equally sized pieces of paper, stirring them up, and drawing one of the slips without looking), that would constitute a **random experiment**. Based on this random experiment, we can define several random variables. For example, let the random variable X be the numerical value showing on a slip that we draw. (We use uppercase letters like X to represent random variables in this primer). The term **random variable** is a bit odd, as it is actually a rule for assigning numerical values to experimental outcomes. In the context of Table P.1, the rule says, "Perform the experiment (stir the slips, and draw one) and for the slip that you obtain assign X to be the number showing." The values that X can take are denoted by corresponding lowercase letters, x, and in this case the values of X are $x = 1, 2, 3,$ or 4.

For the experiment using the population in Table P.1,[1] we can create a number of random variables. Let Y be a discrete random variable designating the color of the slip, with $Y = 1$ denoting

TABLE P.1 The Seussian Slips: A Population

1	2	3	4	4
2	3	3	4	4

[1] A table suitable for classroom experiments can be obtained at www.principlesofeconometrics.com/poe5/extras/table_p1. We thank Veronica Deschner McGregor for the suggestion of "One slip, two slip, white slip, blue slip" for this experiment, inspired by Dr. Seuss's "One Fish Two Fish Red Fish Blue Fish (I Can Read It All by Myself)," Random House Books for Young Readers (1960).

a shaded slip and $Y = 0$ denoting a slip with no shading. The numerical values that Y can take are $y = 0, 1$.

Consider X, the numerical value on the slip. If the slips are equally likely to be chosen after shuffling, then in a large number of experiments (i.e., shuffling and drawing one of the ten slips), 10% of the time we would observe $X = 1$, 20% of the time $X = 2$, 30% of the time $X = 3$, and 40% of the time $X = 4$. These are probabilities that the specific values will occur. We would say, for example, $P(X = 3) = 0.3$. This interpretation is tied to the **relative frequency** of a particular outcome's occurring in a **large** number of experiment replications.

We summarize the probabilities of possible outcomes using a **probability density function** (*pdf*). The *pdf* for a discrete random variable indicates the probability of each possible value occurring. For a discrete random variable X the value of the *pdf* $f(x)$ is the probability that the random variable X takes the value x, $f(x) = P(X = x)$. Because $f(x)$ is a probability, it must be true that $0 \leq f(x) \leq 1$ and, if X takes n possible values $x_1,\ldots, x_n$, then the sum of their probabilities must be one

$$f(x_1) + f(x_2) + \cdots + f(x_n) = 1 \tag{P.1}$$

For discrete random variables, the *pdf* might be presented as a table, such as in Table P.2.

As shown in Figure P.1, the *pdf* may also be represented as a bar graph, with the height of the bar representing the probability with which the corresponding value occurs.

The **cumulative distribution function** (*cdf*) is an alternative way to represent probabilities. The *cdf* of the random variable X, denoted $F(x)$, gives the probability that X is less than or equal to a specific value x. That is,

$$F(x) = P(X \leq x) \tag{P.2}$$

TABLE P.2 **Probability Density Function of X**

x	$f(x)$
1	0.1
2	0.2
3	0.3
4	0.4

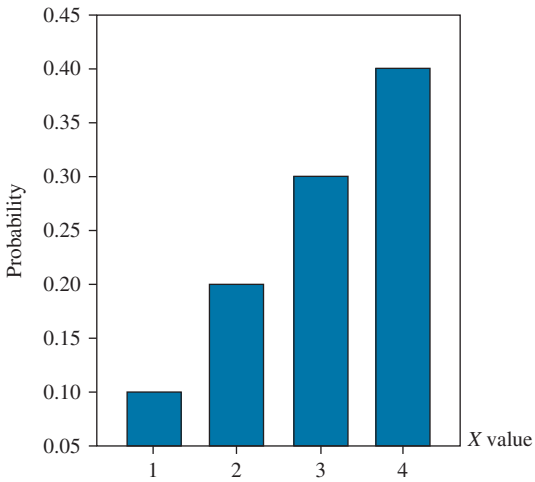


FIGURE P.1 **Probability density function for X.**

EXAMPLE P.1 | Using a *cdf*

Using the probabilities in Table P.2, we find that $F(1) = P(X \leq 1) = 0.1$, $F(2) = P(X \leq 2) = 0.3$, $F(3) = P(X \leq 3) = 0.6$, and $F(4) = P(X \leq 4) = 1$. For example, using the *pdf* $f(x)$ we compute the probability that X is less than or equal to 2 as

$$F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = 0.1 + 0.2 = 0.3$$

Since the sum of the probabilities $P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1$, we can compute the probability that X is greater than 2 as

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - 0.3 = 0.7$$

An important difference between the *pdf* and *cdf* for X is revealed by the question "Using the probability distribution in Table P.2, what is the probability that $X = 2.5$?" This probability is zero because X cannot take this value. The question "What is the probability that X is less than or equal to 2.5?" does have an answer.

$$\begin{aligned} F(2.5) = P(X \leq 2.5) &= P(X = 1) + P(X = 2) \\ &= 0.1 + 0.2 = 0.3 \end{aligned}$$

The cumulative probability can be calculated for any x between $-\infty$ and $+\infty$.

Continuous random variables can take any value in an interval and have an uncountable number of values. Consequently, the probability of any specific value is zero. For continuous random variables, we talk about outcomes being in a certain range. Figure P.2 illustrates the *pdf* $f(x)$ of a continuous random variable X that takes values of x from 0 to infinity. The shape is representative of the distribution for an economic variable such as an individual's income or wage. Areas under the curve represent probabilities that X falls in an interval. The *cdf* $F(x)$ is defined as in (P.2). For this distribution,

$$\begin{aligned} P(10 < X < 20) &= F(20) - F(10) = 0.52236 - 0.17512 \\ &= 0.34724 \end{aligned} \tag{P.3}$$

How are these areas obtained? The integral from calculus gives the area under a curve. We will not compute many integrals in this book.[2] Instead, we will use the computer and compute *cdf* values and probabilities using software commands.

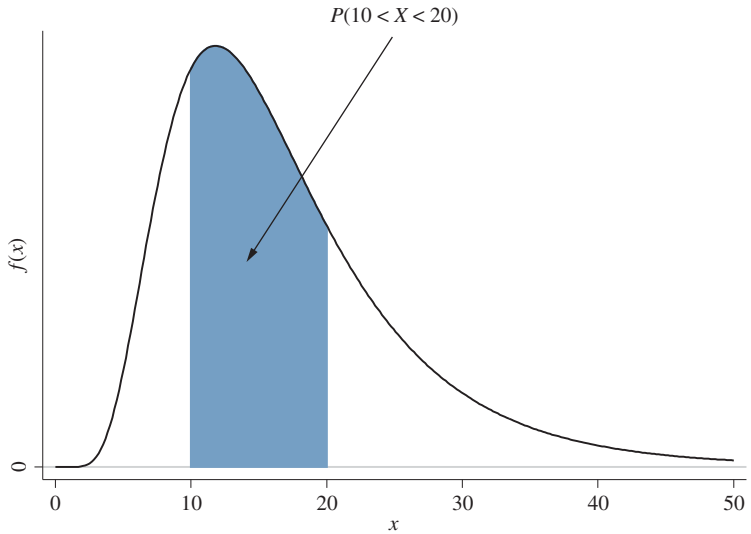


FIGURE P.2 **Probability density function for a continuous random variable.**

[2]See Appendix A.4 for a brief explanation of integrals, and illustrations using integrals to compute probabilities in Appendix B.2.1. The calculations in (P.3) are explained in Appendix B.3.9.

P.3 Joint, Marginal, and Conditional Probabilities

Working with more than one random variable requires a **joint probability density function**. For the population in Table P.1 we defined two random variables, X the numeric value of a randomly drawn slip and the indicator variable Y that equals 1 if the selected slip is shaded, and 0 if it is not shaded.

Using the joint *pdf* for X and Y we can say "The probability of selecting a shaded 2 is 0.10." This is a joint probability because we are talking about the probability of two events occurring simultaneously; the selection takes the value $X = 2$ **and** the slip is shaded so that $Y = 1$. We can write this as

$$P(X = 2 \text{ and } Y = 1) = P(X = 2, Y = 1) = f(x = 2, y = 1) = 0.1$$

The entries in Table P.3 are probabilities $f(x, y) = P(X = x, Y = y)$ of joint outcomes. Like the *pdf* of a single random variable, the sum of the joint probabilities is 1.

P.3.1 Marginal Distributions

Given a joint *pdf*, we can obtain the probability distributions of individual random variables, which are also known as **marginal distributions**. In Table P.3, we see that a shaded slip, $Y = 1$, can be obtained with the values $x = 1$, 2, 3, and 4. The probability that we select a shaded slip is the sum of the probabilities that we obtain a shaded 1, a shaded 2, a shaded 3, and a shaded 4. The probability that $Y = 1$ is

$$P(Y = 1) = f_Y(1) = 0.1 + 0.1 + 0.1 + 0.1 = 0.4$$

This is the sum of the probabilities across the second row of the table. Similarly the probability of drawing a white slip is the sum of the probabilities across the first row of the table, and $P(Y = 0) = f_Y(0) = 0 + 0.1 + 0.2 + 0.3 = 0.6$, where $f_Y(y)$ denotes the *pdf* of the random variable Y. The probabilities $P(X = x)$ are computed similarly by summing down across the values of Y. The joint and marginal distributions are often reported as in Table P.4.[3]

TABLE P.3 **Joint Probability Density Function for X and Y**

	x			
y	1	2	3	4
0	0	0.1	0.2	0.3
1	0.1	0.1	0.1	0.1

TABLE P.4 **Joint and Marginal Probabilities**

y/x	1	2	3	4	$f(y)$
0	0	0.1	0.2	0.3	0.6
1	0.1	0.1	0.1	0.1	0.4
$f(x)$	0.1	0.2	0.3	0.4	1.0

[3] Similar calculations for continuous random variables use integration. See Appendix B.2.3 for an illustration.

P.3.2 Conditional Probability

What is the probability that a randomly chosen slip will take the value 2 **given that** it is shaded? This question is about the **conditional probability** of the outcome $X = 2$ *given* that the outcome $Y = 1$ has occurred. The effect of the conditioning is to reduce the set of possible outcomes. Conditional on $Y = 1$ we only consider the four possible slips that are shaded. One of them is a 2, so the **conditional probability** of the outcome $X = 2$ *given* that $Y = 1$ is 0.25. There is a one in four chance of selecting a 2 given only the shaded slips. Conditioning reduces the size of the population under consideration, and conditional probabilities characterize the reduced population. For discrete random variables the probability that the random variable X takes the value x *given* that $Y = y$ is written $P(X = x|Y = y)$. This conditional probability is given by the **conditional *pdf*** $f(x|y)$

$$f(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f_Y(y)} \tag{P.4}$$

where $f_Y(y)$ is the marginal *pdf* of Y.

EXAMPLE P.2 | Calculating a Conditional Probability

Using the marginal probability $P(Y = 1) = 0.4$, the conditional *pdf* of X given $Y = 1$ is obtained by using (P.4) for each value of X. For example,

$$\begin{aligned} f(x = 2|y = 1) &= P(X = 2|Y = 1) \\ &= \frac{P(X = 2, Y = 1)}{P(Y = 1)} = \frac{f(x = 2, y = 1)}{f_Y(1)} \\ &= \frac{0.1}{0.4} = 0.25 \end{aligned}$$

A key point to remember is that by conditioning we are considering only the subset of a population for which the condition holds. Probability calculations are then based on the "new" population. We can repeat this process for each value of X to obtain the complete conditional *pdf* given in Table P.5.

P.3.3 Statistical Independence

When selecting a shaded slip from Table P.1, the probability of selecting each possible outcome, $x = 1, 2, 3,$ and 4 is 0.25. In the population of shaded slips the numeric values are **equally likely**. The probability of randomly selecting $X = 2$ from the entire population, from the marginal *pdf*, is $P(X = 2) = f_X(2) = 0.2$. This is different from the conditional probability. Knowing that the slip is shaded tells us something about the probability of obtaining $X = 2$. Such random variables are **dependent** in a statistical sense. Two random variables are **statistically independent**, or simply **independent**, if the conditional probability that $X = x$ given that $Y = y$ is the same as the unconditional probability that $X = x$. This means, if X and Y are independent random variables, then

$$P(X = x|Y = y) = P(X = x) \tag{P.5}$$

TABLE P.5 **Conditional Probability of X Given $Y = 1$**

x	1	2	3	4
$f(x\|y = 1)$	0.25	0.25	0.25	0.25

Equivalently, if X and Y are independent, then the conditional *pdf* of X given $Y = y$ is the same as the unconditional, or marginal, *pdf* of X alone.

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = f_X(x) \tag{P.6}$$

Solving (P.6) for the joint *pdf*, we can also say that X and Y are statistically independent if their joint *pdf* factors into the product of their marginal *pdf*s

$$P(X = x, Y = y) = f(x,y) = f_X(x)\, f_Y(y) = P(X = x) \times P(Y = y) \tag{P.7}$$

If (P.5) or (P.7) is true for each and every pair of values *x* and *y*, then *X* and *Y* are statistically independent. This result extends to more than two random variables. The rule allows us to check the independence of random variables X and Y in Table P.4. If (P.7) is violated for any pair of values, then X and Y are not statistically independent. Consider the pair of values $X = 1$ and $Y = 1$.

$$P(X = 1, Y = 1) = f(1,1) = 0.1 \neq f_X(1)\, f_Y(1) = P(X = 1) \times P(Y = 1) = 0.1 \times 0.4 = 0.04$$

The joint probability is 0.1 and the product of the individual probabilities is 0.04. Since these are not equal, we can conclude that X and Y are not statistically independent.

P.4 A Digression: Summation Notation

Throughout this book we will use a **summation sign**, denoted by the symbol $\sum$, to shorten algebraic expressions. Suppose the random variable X takes the values x_1, x_2, …, x_{15}. The sum of these values is $x_1 + x_2 + \cdots + x_{15}$. Rather than write this sum out each time we will represent it as $\sum_{i=1}^{15} x_i$, so that $\sum_{i=1}^{15} x_i = x_1 + x_2 + \cdots + x_{15}$. If we sum n terms, a general number, then the summation will be $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$. In this notation

- The symbol $\sum$ is the capital Greek letter sigma and means "the sum of."
- The letter i is called the **index of summation**. This letter is arbitrary and may also appear as t, j, or k.
- The expression $\sum_{i=1}^{n} x_i$ is read "the sum of the terms x_i, from i equals 1 to n."
- The numbers 1 and n are the **lower limit** and **upper limit** of summation.

The following rules apply to the **summation operation**.

Sum 1. The sum of n values x_1, …, x_n is

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Sum 2. If a is a constant, then

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$$

Sum 3. If a is a constant, then

$$\sum_{i=1}^{n} a = a + a + \cdots + a = na$$

Sum 4. If X and Y are two variables, then

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Sum 5. If X and Y are two variables, then

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

Sum 6. The arithmetic mean (average) of n values of X is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sum 7. A property of the arithmetic mean (average) is that

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Sum 8. We often use an abbreviated form of the summation notation. For example, if $f(x)$ is a function of the values of X,

$$\begin{aligned}\sum_{i=1}^{n} f(x_i) &= f(x_1) + f(x_2) + \cdots + f(x_n)\\ &= \sum_{i} f(x_i) \text{ (“Sum over all values of the index } i\text{”)}\\ &= \sum_{x} f(x) \text{ (“Sum over all possible values of } X\text{”)}\end{aligned}$$

Sum 9. Several summation signs can be used in one expression. Suppose the variable Y takes n values and X takes m values, and let $f(x, y) = x + y$. Then the **double summation** of this function is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_i, y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i + y_j)$$

To evaluate such expressions work from the innermost sum outward. First set $i = 1$ and sum over all values of j, and so on. That is,

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_i, y_j) = \sum_{i=1}^{m}\left[f(x_i, y_1) + f(x_i, y_2) + \cdots + f(x_i, y_n)\right]$$

The *order* of summation does not matter, so

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_i, y_j) = \sum_{j=1}^{n}\sum_{i=1}^{m} f(x_i, y_j)$$

P.5 Properties of Probability Distributions

Figures P.1 and P.2 give us a picture of how frequently values of the random variables will occur. Two key features of a probability distribution are its center (location) and width (dispersion). A key measure of the center is the **mean**, or **expected value**. Measures of dispersion are **variance**, and its square root, the **standard deviation**.

P.5.1 Expected Value of a Random Variable

The **mean** of a random variable is given by its **mathematical expectation**. If X is a discrete random variable taking the values $x_1, \ldots, x_n$, then the mathematical expectation, or **expected value**, of X is

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_n P(X = x_n) \tag{P.8}$$

The expected value, or mean, of X is a weighted average of its values, the weights being the probabilities that the values occur. **The uppercase letter "*E*" represents the expected value operation**. $E(X)$ is read as "the expected value of X." The expected value of X is also called the **mean of *X***. The mean is often symbolized by μ or μ_X. It is the average value of the random variable in an infinite number of repetitions of the underlying experiment. The mean of a random variable is the **population mean**. We use Greek letters for **population parameters** because later on we will use data to estimate these real world unknowns. In particular, keep separate the population mean μ and the arithmetic (or sample) mean $\bar{x}$ that we introduced in Section P.4 as Sum 6. This can be particularly confusing when a conversation includes the term "mean" without the qualifying term "population" or "arithmetic." Pay attention to the usage context.

EXAMPLE P.3 | Calculating an Expected Value

For the population in Table P.1, the expected value of X is

$$\begin{aligned} E(X) &= 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + 4 \times P(X = 4) \\ &= (1 \times 0.1) + (2 \times 0.2) + (3 \times 0.3) + (4 \times 0.4) = 3 \end{aligned}$$

For a discrete random variable the probability that X takes the value x is given by its *pdf* $f(x)$, $P(X = x) = f(x)$. The expected value in (P.8) can be written equivalently as

$$\begin{aligned} \mu_X = E(X) &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n) \\ &= \sum_{i=1}^{n} x_i f(x_i) = \sum_{x} x f(x) \end{aligned} \tag{P.9}$$

Using (P.9), the expected value of X, the numeric value on a randomly drawn slip from Table P.1 is

$$\mu_X = E(X) = \sum_{x=1}^{4} x f(x) = (1 \times 0.1) + (2 \times 0.2) + (3 \times 0.3) + (4 \times 0.4) = 3$$

What does this mean? Draw one "slip" at random from Table P.1, and observe its numerical value X. This constitutes an experiment. If we repeat this experiment many times, the values $x = 1$, 2, 3, and 4 will appear 10%, 20%, 30%, and 40% of the time, respectively. The arithmetic average of all the numerical values will approach $\mu_X = 3$, as the number of experiments becomes large. The key point is that **the expected value of the random variable is the average value that occurs in many repeated trials of an experiment**.

For continuous random variables, the interpretation of the expected value of X is unchanged—it is the average value of X if many values are obtained by repeatedly performing the underlying random experiment.[4]

[4] Since there are now an uncountable number of values to sum, mathematically we must replace the "summation over all possible values" in (P.9) by the "integral over all possible values." See Appendix B.2.2 for a brief discussion.

P.5.2 Conditional Expectation

Many economic questions are formulated in terms of **conditional expectation**, or the **conditional mean**. One example is "What is the mean (expected value) wage of a person who has 16 years of education?" In expected value notation, what is $E(WAGE|EDUCATION = 16)$? For a discrete random variable, the calculation of conditional expected value uses (P.9) with the conditional *pdf* $f(x|y)$ replacing $f(x)$, so that

$$\mu_{X|Y} = E(X|Y = y) = \sum_x x f(x|y)$$

EXAMPLE P.4 | Calculating a Conditional Expectation

Using the population in Table P.1, what is the expected numerical value of X given that $Y = 1$, the slip is shaded? The conditional probabilities $f(x|y = 1)$ are given in Table P.5. The conditional expectation of X is

$$\begin{aligned} E(X|Y = 1) &= \sum_{x=1}^{4} x f(x|1) = 1 \times f(1|1) + 2 \times f(2|1) \\ &\quad + 3 \times f(3|1) + 4 \times f(4|1) \\ &= 1(0.25) + 2(0.25) + 3(0.25) + 4(0.25) = 2.5 \end{aligned}$$

The average value of X in many repeated trials of the experiment of drawing from the shaded slips is 2.5. This example makes a good point about expected values in general, namely that the expected value of X does not have to be a value that X can take. The expected value of X is **not** the value that you expect to occur in any single experiment.

What is the conditional expectation of X given that $Y = y$ if the random variables are statistically independent? If X and Y are statistically independent the conditional *pdf* $f(x|y)$ equals the *pdf* of X alone, $f(x)$, as shown in (P.6). The conditional expectation is then

$$E(X|Y = y) = \sum_x x f(x|y) = \sum_x x f(x) = E(X)$$

If X and Y are statistically independent, conditioning does not affect the expected value.

P.5.3 Rules for Expected Values

Functions of random variables are also random. If $g(X)$ is a function of the random variable X, such as $g(X) = X^2$, then $g(X)$ is also random. If X is a discrete random variable, then the expected value of $g(X)$ is obtained using calculations similar to those in (P.9).

$$E\left[g(X)\right] = \sum_x g(x) f(x) \tag{P.10}$$

For example, if a is a constant, then $g(X) = aX$ is a function of X, and

$$\begin{aligned} E(aX) = E\left[g(X)\right] &= \sum_x g(x) f(x) \\ &= \sum_x ax f(x) = a\sum_x x f(x) \\ &= aE(X) \end{aligned}$$

Similarly, if a and b are constants, then we can show that

$$E(aX + b) = aE(X) + b \tag{P.11}$$

If $g_1(X)$ and $g_2(X)$ are functions of X, then

$$E\left[g_1(X) + g_2(X)\right] = E\left[g_1(X)\right] + E\left[g_2(X)\right] \tag{P.12}$$

This rule extends to any number of functions. Remember the phrase "**the expected value of a sum is the sum of the expected values.**"

P.5.4 Variance of a Random Variable

The **variance** of a discrete or continuous random variable X is the expected value of

$$g(X) = \left[X - E(X)\right]^2$$

The variance of a random variable is important in characterizing the scale of measurement and the spread of the probability distribution. We give it the symbol σ^2, or σ_X^2, read "sigma squared." The variance σ^2 has a Greek symbol because it is a population parameter. Algebraically, letting $E(X) = \mu$, using the rules of expected values and the fact that $E(X) = \mu$ is not random, we have

$$\begin{aligned} \text{var}(X) = \sigma_X^2 &= E(X-\mu)^2 \\ &= E\left(X^2 - 2\mu X + \mu^2\right) = E\left(X^2\right) - 2\mu E(X) + \mu^2 \\ &= E\left(X^2\right) - \mu^2 \end{aligned} \tag{P.13}$$

We use the letters "var" to represent variance, and var(X) is read as "**the variance of X**," where X is a random variable. The calculation $\text{var}(X) = E\left(X^2\right) - \mu^2$ is usually simpler than $\text{var}(X) = E(X-\mu)^2$, but the solution is the same.

EXAMPLE P.5 | Calculating a Variance

For the population in Table P.1, we have shown that $E(X) = \mu = 3$. Using (P.10), the expectation of the random variable $g(X) = X^2$ is

$$\begin{aligned} E\left(X^2\right) &= \sum_{x=1}^{4} g(x)\, f(x) = \sum_{x=1}^{4} x^2 f(x) \\ &= \left[1^2 \times 0.1\right] + \left[2^2 \times 0.2\right] + \left[3^2 \times 0.3\right] + \left[4^2 \times 0.4\right] = 10 \end{aligned}$$

Then, the variance of the random variable X is

$$\text{var}(X) = \sigma_X^2 = E\left(X^2\right) - \mu^2 = 10 - 3^2 = 1$$

The square root of the variance is called the **standard deviation**; it is denoted by σ or sometimes as σ_X if more than one random variable is being discussed. It also measures the spread or dispersion of a probability distribution and has the advantage of being in the same units of measure as the random variable.

A useful property of variances is the following. Let a and b be constants, then

$$\text{var}(aX + b) = a^2 \text{var}(X) \tag{P.14}$$

An additive constant like b changes the mean (expected value) of a random variable, but it does not affect its dispersion (variance). A multiplicative constant like a affects the mean, and it affects the variance by the **square** of the constant.

To see this, let $Y = aX + b$. Using (P.11)

$$E(Y) = \mu_Y = aE(X) + b = a\mu_X + b$$

Then

$$\begin{aligned} \text{var}(aX+b) = \text{var}(Y) &= E\left[\left(Y - \mu_Y\right)^2\right] = E\left[\left(aX + b - \left(a\mu_X + b\right)\right)^2\right] \\ &= E\left[\left(aX - a\mu_X\right)^2\right] = E\left[a^2\left(X - \mu_X\right)^2\right] \\ &= a^2 E\left[\left(X - \mu_X\right)^2\right] = a^2 \text{var}(X) \end{aligned}$$

The variance of a random variable is the *average* squared difference between the random variable X and its mean value μ_X. The larger the variance of a random variable, the more "spread out" the

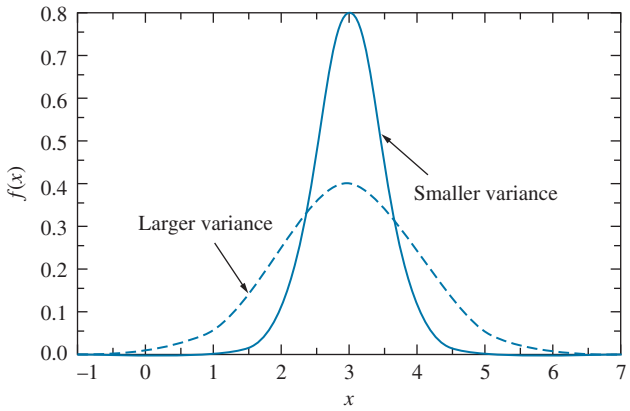


FIGURE P.3 **Distributions with different variances.**

values of the random variable are. Figure P.3 shows two *pdf*s for a continuous random variable, both with mean $\mu = 3$. The distribution with the smaller variance (the solid curve) is less spread out about its mean.

P.5.5 Expected Values of Several Random Variables

Let X and Y be random variables. The rule "the expected value of the sum is the sum of the expected values" applies. Then[5]

$$E(X + Y) = E(X) + E(Y) \tag{P.15}$$

Similarly

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{P.16}$$

The product of random variables is not as easy. $E(XY) = E(X)E(Y)$ **if** X and Y are independent. These rules can be extended to more random variables.

P.5.6 Covariance Between Two Random Variables

The **covariance** between X and Y is a measure of linear association between them. Think about two continuous variables, such as height and weight of children. We expect that there is an association between height and weight, with taller than average children tending to weigh more than the average. The product of X minus its mean times Y minus its mean is

$$\left(X - \mu_X\right)\left(Y - \mu_Y\right) \tag{P.17}$$

In Figure P.4, we plot values (x and y) of X and Y that have been constructed so that $E(X) = E(Y) = 0$.

The x and y values of X and Y fall predominately in quadrants I and III, so that the arithmetic average of the values $(x - \mu_X)(y - \mu_Y)$ is positive. We define the covariance between two random variables as the expected (population average) value of the product in (P.17).

$$\text{cov}(X, Y) = \sigma_{XY} = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] = E(XY) - \mu_X\mu_Y \tag{P.18}$$

We use the letters "cov" to represent covariance, and cov(X, Y) is read as "**the covariance between X and Y**," where X and Y are random variables. The covariance σ_{XY} of the random variables underlying Figure P.4 is positive, which tells us that when the values x are greater

[5]These results are proven in Appendix B.1.4.

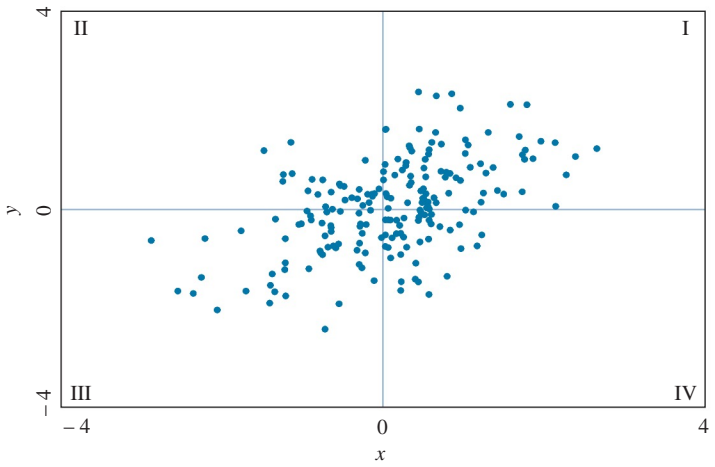


FIGURE P.4 **Correlated data.**

than μ_X, then the values y also tend to be greater than μ_Y; and when the values x are below μ_X, then the values y also tend to be less than μ_Y. If the random variables values tend primarily to fall in quadrants II and IV, then $(x-\mu_X)(y-\mu_Y)$ will tend to be negative and σ_{XY} will be negative. If the random variables values are spread evenly across the four quadrants, and show neither positive nor negative association, then the covariance is zero. The sign of σ_{XY} tells us whether the two random variables X and Y are positively associated or negatively associated.

Interpreting the actual value of σ_{XY} is difficult because X and Y may have different units of measurement. Scaling the covariance by the standard deviations of the variables eliminates the units of measurement, and defines the **correlation** between X and Y

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{P.19}$$

As with the covariance, the correlation ρ between two random variables measures the degree of *linear* association between them. However, unlike the covariance, the correlation must lie between –1 and 1. Thus, the correlation between X and Y is 1 or –1 if X is a perfect positive or negative linear function of Y. If there is *no linear* association between X and Y, then $\text{cov}(X, Y) = 0$ and $\rho = 0$. For other values of correlation the magnitude of the absolute value $|\rho|$ indicates the "strength" of the linear association between the values of the random variables. In Figure P.4, the correlation between X and Y is $\rho = 0.5$.

EXAMPLE P.6 | Calculating a Correlation

To illustrate the calculation, reconsider the population in Table P.1 with joint *pdf* given in Table P.4. The expected value of XY is

$$\begin{aligned} E(XY) &= \sum_{y=0}^{1}\sum_{x=1}^{4} xy f(x,y) \\ &= (1\times 0\times 0) + (2\times 0\times 0.1) + (3\times 0\times 0.2) \\ &\quad + (4\times 0\times 0.3) + (1\times 1\times 0.1) \\ &\quad + (2\times 1\times 0.1) + (3\times 1\times 0.1) + (4\times 1\times 0.1) \\ &= 0.1 + 0.2 + 0.3 + 0.4 \\ &= 1 \end{aligned}$$

The random variable X has expected value $E(X) = \mu_X = 3$ and the random variable Y has expected value $E(Y) = \mu_Y = 0.4$. Then the covariance between X and Y is

$$\text{cov}(X,Y) = \sigma_{XY} = E(XY) - \mu_X\mu_Y = 1 - 3\times(0.4) = -0.2$$

The correlation between X and Y is

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{-0.2}{\sqrt{1}\times\sqrt{0.24}} = -0.4082$$

If X and Y are independent random variables, then their covariance and correlation are zero. The converse of this relationship is **not** true. Independent random variables X and Y have zero covariance, indicating that there is no linear association between them. However, just because the covariance or correlation between two random variables is zero **does not** mean that they are necessarily independent. There may be more complicated nonlinear associations such as $X^2 + Y^2 = 1$.

In (P.15) we obtain the expected value of a sum of random variables. There are similar rules for variances. If a and b are constants, then

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\ \text{cov}(X, Y) \tag{P.20}$$

A significant point to note is that the variance of a sum is **not** just the sum of the variances. There is a covariance term present. Two special cases of (P.20) are

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \tag{P.21}$$

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y) \tag{P.22}$$

To show that (P.22) is true, let $Z = X - Y$. Using the rules of expected value

$$E(Z) = \mu_Z = E(X) - E(Y) = \mu_X - \mu_Y$$

The variance of $Z = X - Y$ is obtained using the basic definition of variance, with some substitution,

$$\begin{aligned}
\text{var}(X - Y) = \text{var}(Z) &= E\left[(Z - \mu_Z)^2\right] = E\left[\left(X - Y - (\mu_X - \mu_Y)\right)^2\right] \\
&= E\left\{\left[(X - \mu_X) - (Y - \mu_Y)\right]^2\right\} \\
&= E\left\{(X - \mu_X)^2 + (Y - \mu_Y)^2 - 2(X - \mu_X)(Y - \mu_Y)\right\} \\
&= E\left[(X - \mu_X)^2\right] + E\left[(Y - \mu_Y)^2\right] - 2E\left[(X - \mu_X)(Y - \mu_Y)\right] \\
&= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)
\end{aligned}$$

If X and Y are independent, or if $\text{cov}(X, Y) = 0$, then

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) \tag{P.23}$$

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \tag{P.24}$$

These rules extend to more random variables.

P.6 Conditioning

In Table P.4, we summarized the joint and marginal probability functions for the random variables X and Y defined on the population in Table P.1. In Table P.6 we make two modifications. First, the probabilities are expressed as fractions. The many calculations below are simpler using arithmetic

TABLE P.6 **Joint, Marginal, and Conditional Probabilities**

y/x	1	2	3	4	$f(y)$	$f(y\|x=1)$	$f(y\|x=2)$	$f(y\|x=3)$	$f(y\|x=4)$
0	0	1/10	2/10	3/10	6/10	0	1/2	2/3	3/4
1	1/10	1/10	1/10	1/10	4/10	1	1/2	1/3	1/4
$f(x)$	1/10	2/10	3/10	4/10					
$f(x\|y=0)$	0	1/6	2/6	3/6					
$f(x\|y=1)$	1/4	1/4	1/4	1/4					

with fractions. Second, we added the conditional probability functions (P.4) for Y given each of the values that X can take and the conditional probability functions for X given each of the values that Y can take. Now would be a good time for you to review Section P.3.2 on conditional probability. For example, what is the probability that $Y = 0$ given that $X = 2$? That is, if we only consider population members with $X = 2$, what is the probability that $Y = 0$? There are only two population elements with $X = 2$, one with $Y = 0$ and one with $Y = 1$. The probability of randomly selecting $Y = 0$ is one-half. For discrete random variables, the conditional probability is calculated as the joint probability divided by the probability of the conditioning event.

$$f(y = 0|x = 2) = P(Y = 0|X = 2) = \frac{P(X = 2, Y = 0)}{P(X = 2)} = \frac{1/10}{2/10} = \frac{1}{2}$$

In the following sections, we discuss the concepts of **conditional expectation** and **conditional variance**.

P.6.1 Conditional Expectation

Many economic questions are formulated in terms of a **conditional expectation**, or the **conditional mean**. One example is "What is the mean wage of a person who has 16 years of education?" In expected value notation, what is $E(WAGE|EDUC = 16)$? The effect of conditioning on the value of $EDUC$ is to reduce the population of interest to only individuals with 16 years of education. The mean, or expected value, of wage for these individuals may be quite different than the mean wage for all individuals regardless of years of education, $E(WAGE)$, which is the **unconditional expectation** or **unconditional mean**.

For discrete random variables,[6] the calculation of a conditional expected value uses equation (P.9) with the conditional *pdf* replacing the usual *pdf*, so that

$$\begin{aligned} E(X|Y = y) &= \sum_x x f(x|y) = \sum_x x P(X = x|Y = y) \\ E(Y|X = x) &= \sum_y y f(y|x) = \sum_y y P(Y = y|X = x) \end{aligned} \tag{P.25}$$

EXAMPLE P.7 | Conditional Expectation

Using the population in Table P.1, what is the expected numerical value of X given that $Y = 1$? The conditional probabilities $P(X = x|Y = 1) = f(x|y = 1) = f(x|1)$ are given in Table P.6. The conditional expectation of X is

$$\begin{aligned} E(X|Y = 1) &= \sum_{x=1}^{4} x f(x|1) \\ &= [1 \times f(1|1)] + [2 \times f(2|1)] + [3 \times f(3|1)] \\ &\quad + [4 \times f(4|1)] \\ &= 1(1/4) + 2(1/4) + 3(1/4) + 4(1/4) = 10/4 \\ &= 5/2 \end{aligned}$$

The average value of X in many repeated trials of the experiment of drawing from the shaded slips ($Y = 1$) is 2.5. This example makes a good point about expected values in general, namely that the expected value of X does not have to be a value that X can take. The expected value of X is **not** the value that you expect to occur in any single experiment. It is the average value of X after many repetitions of the experiment.

What is the expected value of X given that we only consider values where $Y = 0$? Confirm that $E(X|Y = 0) = 10/3$. For comparison purposes recall from Section P.5.1 that the **unconditional expectation** of X is $E(X) = 3$.

Similarly, if we condition on the X values, the conditional expectations of Y are

$$\begin{aligned} E(Y|X = 1) &= \sum_y y f(y|1) = 0(0) + 1(1) = 1 \\ E(Y|X = 2) &= \sum_y y f(y|2) = 0(1/2) + 1(1/2) = 1/2 \\ E(Y|X = 3) &= \sum_y y f(y|3) = 0(2/3) + 1(1/3) = 1/3 \\ E(Y|X = 4) &= \sum_y y f(y|4) = 0(3/4) + 1(1/4) = 1/4 \end{aligned}$$

Note that $E(Y|X)$ varies as X varies; it is a function of X. For comparison, the unconditional expectation of Y, $E(Y)$, is

$$E(Y) = \sum_y y f(y) = 0(6/10) + 1(4/10) = 2/5$$

[6] For continuous random variables the sums are replaced by integrals. See Appendix B.2.

P.6.2 Conditional Variance

The **unconditional variance** of a discrete random variable X is

$$\text{var}(X) = \sigma_X^2 = E\left[\left(X - \mu_X\right)^2\right] = \sum_x \left(x - \mu_X\right)^2 f(x) \tag{P.26}$$

It measures how much variation there is in X around the unconditional mean of X, μ_X. For example, the unconditional variance var(*WAGE*) measures the variation in *WAGE* around the unconditional mean $E(WAGE)$. In (P.13) we show that equivalently

$$\text{var}(X) = \sigma_X^2 = E\left(X^2\right) - \mu_X^2 = \sum_x x^2 f(x) - \mu_X^2 \tag{P.27}$$

In Section P.6.1 we discussed how to answer the question "What is the mean wage of a person who has 16 years of education?" Now we ask "How much variation is there in wages for a person who has 16 years of education?" The answer to this question is given by the **conditional variance**, var $(WAGE|EDUC = 16)$. The conditional variance measures the variation in *WAGE* around the conditional mean $E(WAGE|EDUC = 16)$ for individuals with 16 years of education. The conditional variance of *WAGE* for individuals with 16 years of education is the average squared difference in the population between *WAGE* and the conditional mean of *WAGE*,

$$\underbrace{\text{var}(WAGE \mid EDUC = 16)}_{\text{conditional variance}} = E\left\{\left[WAGE - \underbrace{E(WAGE \mid EDUC = 16)}_{\text{conditional mean}}\right]^2 \middle| EDUC = 16\right\}$$

To obtain the conditional variance we modify the definitions of variance in equations (P.26) and (P.27); replace the unconditional mean $E(X) = \mu_X$ with the conditional mean $E(X|Y = y)$, and the unconditional *pdf* $f(x)$ with the conditional *pdf* $f(x|y)$. Then

$$\text{var}(X|Y = y) = E\left\{\left[X - E(X|Y = y)\right]^2 \middle| Y = y\right\} = \sum_x \left(x - E(X|Y = y)\right)^2 f(x|y) \tag{P.28}$$

or

$$\text{var}(X|Y = y) = E\left(X^2|Y = y\right) - \left[E(X|Y = y)\right]^2 = \sum_x x^2 f(x|y) - \left[E(X|Y = y)\right]^2 \tag{P.29}$$

EXAMPLE P.8 | Conditional Variance

For the population in Table P.1, the unconditional variance of X is $\text{var}(X) = 1$. What is the variance of X given that $Y = 1$? To use (P.29) first compute

$$\begin{aligned} &E\left(X^2|Y = 1\right) \\ &= \sum_x x^2 f(x|Y = 1) \\ &= 1^2(1/4) + 2^2(1/4) + 3^2(1/4) + 4^2(1/4) = 15/2 \end{aligned}$$

Then

$$\begin{aligned} \text{var}(X|Y = 1) &= E\left(X^2|Y = 1\right) - \left[E(X|Y = 1)\right]^2 \\ &= 15/2 - (5/2)^2 = 5/4 \end{aligned}$$

In this case, the conditional variance of X, given that $Y = 1$, is larger than the unconditional variance of X, $\text{var}(X) = 1$.

To calculate the conditional variance of X given that $Y = 0$, we first obtain

$$\begin{aligned} E\left(X^2|Y = 0\right) &= \sum_x x^2 f(x|Y = 0) \\ &= 1^2(0) + 2^2(1/6) + 3^2(2/6) + 4^2(3/6) \\ &= 35/3 \end{aligned}$$

Then

$$\begin{aligned} \text{var}(X|Y = 0) &= E\left(X^2|Y = 0\right) - \left[E(X|Y = 0)\right]^2 \\ &= 35/3 - (10/3)^2 = 5/9 \end{aligned}$$

In this case, the conditional variance of X, given that $Y = 0$, is smaller than the unconditional variance of X, $\text{var}(X) = 1$. These examples have illustrated that in general the conditional variance can be larger or smaller than the unconditional variance. Try working out $\text{var}(Y|X = 1)$, $\text{var}(Y|X = 2)$, $\text{var}(Y|X = 3)$, and $\text{var}(Y|X = 4)$.

P.6.3 Iterated Expectations

The Law of **Iterated Expectations** says that we can find the expected value of Y in two steps. First, find the conditional expectation $E(Y|X)$. Second, find the expected value $E(Y|X)$ treating X as random.

$$\text{Law of Iterated Expectations: } E(Y) = E_X\left[E(Y|X)\right] = \sum_x E(Y|X=x)\, f_X(x) \tag{P.30}$$

In this expression we put an "X" subscript in the expectation $E_X[E(Y|X)]$ and the probability function $f_X(x)$ to emphasize that we are treating X as random. The Law of Iterated Expectations is true for both discrete and continuous random variables.

EXAMPLE P.9 | Iterated Expectation

Consider the conditional expectation $E(X|Y=y) = \sum_x xf(x|y)$. As we computed in Section P.6.1, $E(X|Y=0) = 10/3$ and $E(X|Y=1) = 5/2$. Similarly, the conditional expectation $E(Y|X=x) = \sum_y yf(y|x)$. For the population in Table P.1, these conditional expectations were calculated in Section P.6.1 to be $E(Y|X=1) = 1$, $E(Y|X=2) = 1/2$, $E(Y|X=3) = 1/3$ and $E(Y|X=4) = 1/4$. Note that $E(Y|X=x)$ changes when x changes. If X is allowed to vary randomly[7] then the conditional expectation varies randomly. The conditional expectation is a function of X, or $E(Y|X) = g(X)$, and is random when viewed this way. Using (P.10) we can find the expected value of $g(X)$.

$$\begin{aligned}
E_X\left[E(Y|X)\right] = E_X\left[g(X)\right] &= \sum_x g(x) f_X(x) = \sum_x E(Y|X=x)\, f_X(x) \\
&= \left[E(Y|X=1)\, f_X(1)\right] + \left[E(Y|X=2)\, f_X(2)\right] \\
&\quad + \left[E(Y|X=3)\, f_X(3)\right] + \left[E(Y|X=4)\, f_X(4)\right] \\
&= 1(1/10) + (1/2)(2/10) + (1/3)(3/10) \\
&\quad + (1/4)(4/10) = 2/5
\end{aligned}$$

If we draw many values x from the population in Table P.1, the average of $E(Y|X)$ is 2/5. For comparison the "unconditional" expectation of Y is $E(Y) = 2/5$. $E_X[E(Y|X)]$ and $E(Y)$ are the same.

Proof of the Law of Iterated Expectations To prove the Law of Iterated Expectations we make use of relationships between joint, marginal, and conditional *pdf*s that we introduced in Section P.3. In Section P.3.1 we discussed *marginal distributions*. Given a joint *pdf* $f(x, y)$ we can obtain the *marginal pdf* of y alone $f_Y(y)$ by summing, for each y, the joint *pdf* $f(x, y)$ across all values of the variable we wish to eliminate, in this case x. That is, for Y and X,

$$\begin{aligned}
f(y) = f_Y(y) &= \sum_x f(x, y) \\
f(x) = f_X(x) &= \sum_y f(x, y)
\end{aligned} \tag{P.31}$$

Because $f(\,)$ is used to represent *pdf*s in general, sometimes we will put a subscript, X or Y, to be very clear about which variable is random.

Using equation (P.4) we can define the conditional *pdf* of y given $X = x$ as

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

Rearrange this expression to obtain

$$f(x, y) = f(y|x)\, f_X(x) \tag{P.32}$$

A joint *pdf* is the product of the conditional *pdf* and the *pdf* of the conditioning variable.

[7] Imagine shuffling the population elements and randomly choosing one. This is an experiment and the resulting number showing is a value of X. By doing this repeatedly X varies randomly.

To show that the Law of Iterated Expectations is true[8] we begin with the definition of the expected value of Y, and operate with the summation.

$$\begin{aligned} E(Y) &= \sum_y y f(y) = \sum_y y \left[\sum_x f(x,y)\right] && \text{[substitute for } f(y)\text{]} \\ &= \sum_y y \left[\sum_x f(y|x)\, f_X(x)\right] && \text{[substitute for } f(x,y)\text{]} \\ &= \sum_x \left[\sum_y y f(y|x)\right] f_X(x) && \text{[change order of summation]} \\ &= \sum_x E(Y|x) f_X(x) && \text{[recognize the conditional expectation]} \\ &= E_X\left[E(Y|X)\right] \end{aligned}$$

While this result may seem an esoteric oddity it is very important and widely used in modern econometrics.

P.6.4 Variance Decomposition

Just as we can break up the expected value using the Law of Iterated Expectations we can decompose the variance of a random variable into two parts.

$$\text{Variance Decomposition: } \operatorname{var}(Y) = \operatorname{var}_X\left[E(Y|X)\right] + E_X\left[\operatorname{var}(Y|X)\right] \tag{P.33}$$

This "beautiful" result[9] says that the variance of the random variable Y equals the sum of the variance of the conditional mean of Y given X and the mean of the conditional variance of Y given X. In this section we will discuss this result.[10]

Suppose that we are interested in the wages of the population consisting of working adults. How much variation do wages display in the population? If *WAGE* is the wage of a randomly drawn population member, then we are asking about the variance of *WAGE*, that is, var(*WAGE*). The variance decomposition says

$$\operatorname{var}(WAGE) = \operatorname{var}_{EDUC}\left[E(WAGE|EDUC)\right] + E_{EDUC}\left[\operatorname{var}(WAGE|EDUC)\right]$$

$E(WAGE|EDUC)$ is the expected value of *WAGE* given a specific value of education, such as $EDUC = 12$ or $EDUC = 16$. $E(WAGE|EDUC = 12)$ is the average *WAGE* in the population, given that we only consider workers who have 12 years of education. If *EDUC* changes then the conditional mean $E(WAGE|EDUC)$ changes, so that $E(WAGE|EDUC = 16)$ is not the same as $E(WAGE|EDUC = 12)$, and in fact we expect $E(WAGE|EDUC = 16) > E(WAGE|EDUC = 12)$; more education means more "human capital" and thus the average wage should be higher. The first component in the variance decomposition $\operatorname{var}_{EDUC}\left[E(WAGE|EDUC)\right]$ measures the variation in $E(WAGE|EDUC)$ due to variation in education.

The second part of the variance decomposition is $E_{EDUC}\left[\operatorname{var}(WAGE|EDUC)\right]$. If we restrict our attention to population members who have 12 years of education, the mean wage is $E(WAGE|EDUC = 12)$. Within the group of workers who have 12 years of education we will observe wide ranges of wages. For example, using one sample of *CPS* data from 2013,[11] wages for those with 12 years of education varied from \$3.11/hour to \$100.00/hour; for those with 16 years of education wages varied from \$2.75/hour to \$221.10/hour. For workers with 12 and 16 years of education that variation is measured by $\operatorname{var}(WAGE|EDUC = 12)$ and

[8] The proof for continuous variables is in Appendix B.2.4.

[9] Tony O'Hagan, "A Thing of Beauty," *Significance Magazine*, Volume 9 Issue 3 (June 2012), 26–28.

[10] The proof of the variance decomposition is given in Appendix B.1.8 and Example B.1.

[11] The data file *cps5*.

var($WAGE|EDUC = 16$). The term $E_{EDUC}\left[\text{var}(WAGE|EDUC)\right]$ measures the average of var($WAGE|EDUC$) as education changes.

To summarize, the variation of *WAGE* in the population can be attributed to two sources: variation in the conditional mean $E(WAGE|EDUC)$ and variation due to changes in education in the conditional variance of *WAGE* given education.

P.6.5 Covariance Decomposition

Recall that the covariance between two random variables Y and X is $\text{cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$. For discrete random variables this is

$$\text{cov}(X,Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)\ f(x,y)$$

By using the relationships between marginal, conditional and joint *pdf*s we can show

$$\text{cov}(X,Y) = \sum_x (x-\mu_X)\, E(Y|X=x)\, f(x) \tag{P.34}$$

Recall that $E(Y|X) = g(X)$ so this result says that the covariance between X and Y can be calculated as the expected value of X, minus its mean, times a function of X, $\text{cov}(X,Y) = E_X[(X-\mu_X)E(Y|X)]$.

An important special case is important in later chapters. When the conditional expectation of Y given X is a constant, $E(Y|X=x) = c$, then

$$\text{cov}(X,Y) = \sum_x (x-\mu_X)\, E(Y|X=x)\, f(x) = c\sum_x (x-\mu_X)\, f(x) = 0$$

A special case is $E(Y|X=x) = 0$, which by direct substitution implies $\text{cov}(X,Y) = 0$.

EXAMPLE P.10 | Covariance Decomposition

To illustrate we compute $\text{cov}(X,Y)$ for the population in Table P.1 using the covariance decomposition. We have computed that $\text{cov}(X,Y) = -0.2$ in Section P.5.6. The ingredients are the values of the random variable X, its mean $\mu_X = 3$, the probabilities $P(X=x) = f(x)$ and conditional expectations

$$E(Y|X=1) = 1,\ E(Y|X=2) = 1/2,$$
$$E(Y|X=3) = 1/3 \text{ and } E(Y|X=4) = 1/4$$

Using the covariance decomposition we have

$$\begin{aligned}\text{cov}(X,Y) &= \sum_x (x-\mu_X)\, E(Y|X=x)\, f(x)\\ &= (1-3)(1)(1/10) + (2-3)(1/2)(2/10)\\ &\quad + (3-3)(1/3)(3/10) + (4-3)(1/4)(4/10)\\ &= -2/10 - 1/10 + 1/10 = -2/10 = -0.2\end{aligned}$$

We see that the covariance decomposition yields the correct result, and it is convenient in this example.

P.7 The Normal Distribution

In the previous sections we discussed random variables and their *pdf*s in a general way. In real economic contexts, some specific *pdf*s have been found to be very useful. The most important is the **normal distribution**. If X is a normally distributed random variable with mean μ and variance σ^2, it is symbolized as $X \sim N(\mu, \sigma^2)$. The *pdf* of X is given by the impressive formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \tag{P.35}$$

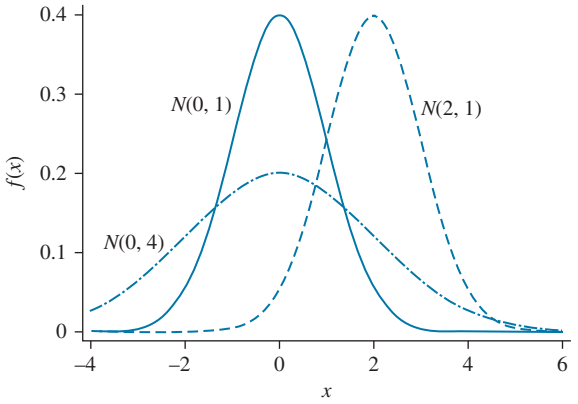


FIGURE P.5 **Normal probability density functions $N(\mu, \sigma^2)$.**

where exp(a) denotes the exponential[12] function e^a. The mean μ and variance σ^2 are the parameters of this distribution and determine its center and dispersion. The range of the continuous normal random variable is from minus infinity to plus infinity. Pictures of the normal *pdf*s are given in Figure P.5 for several values of the mean and variance. Note that the distribution is symmetric and centered at μ.

Like all continuous random variables, probabilities involving normal random variables are found as areas under the *pdf*. For calculating probabilities both computer software and statistical tables values make use of the relation between a normal random variable and its "standardized" equivalent. A **standard normal random variable** is one that has a normal *pdf* with mean 0 and variance 1. If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \tag{P.36}$$

The standard normal random variable Z is so widely used that its *pdf* and *cdf* are given their own special notation. The *cdf* is denoted $\Phi(z) = P(Z \leq z)$. Computer programs, and Statistical Table 1 in Appendix D give values of $\Phi(z)$. The *pdf* for the standard normal random variable is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad -\infty < z < \infty$$

Values of the density function are given in Statistical Table 6 in Appendix D. To calculate normal probabilities, remember that the distribution is symmetric, so that $P(Z > a) = P(Z < -a)$, and $P(Z > a) = P(Z \geq a)$, since the probability of any one point is zero for a continuous random variable. If $X \sim N(\mu, \sigma^2)$ and a and b are constants, then

$$P(X \leq a) = P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = P\left(Z \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right) \tag{P.37}$$

$$P(X > a) = P\left(\frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right) = P\left(Z > \frac{a - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) \tag{P.38}$$

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \tag{P.39}$$

[12] See Appendix A.1.2 for a review of exponents.

EXAMPLE P.11 | Normal Distribution Probability Calculation

For example, if $X \sim N(3, 9)$, then

$P(4 \le X \le 6) = P(0.33 \le Z \le 1) = \Phi(1) - \Phi(0.33) = 0.8413 - 0.6293 = 0.2120$

In addition to finding normal probabilities we sometimes must find a value z_α of a standard normal random variable such that $P(Z \le z_\alpha) = \alpha$. The value z_α is called the **100α-percentile**. For example, $z_{0.975}$ is the value of Z such that $P(Z \le z_{0.975}) = 0.975$. This particular percentile can be found using Statistical Table 1, Cumulative Probabilities for the Standard Normal Distribution. The cumulative probability associated with the value $z = 1.96$ is $P(Z \le 1.96) = 0.975$, so that the 97.5 percentile is $z_{0.975} = 1.96$. Using Statistical Table 1 we can only roughly obtain other percentiles. Using the cumulative probabilities $P(Z \le 1.64) = 0.9495$ and $P(Z \le 1.65) = 0.9505$ we can say that the 95th percentile of the **standard normal distribution** is between 1.64 and 1.65, and is about 1.645.

Luckily computer software makes these approximations unnecessary. The **inverse normal** function finds percentiles z_α given α. Formally, if $P(Z \le z_\alpha) = \Phi(z_\alpha) = \alpha$ then $z_\alpha = \Phi^{-1}(\alpha)$. Econometric software, even spreadsheets, have the inverse normal function built in. Some commonly used percentiles are shown in Table P.7. In the last column are the percentiles rounded to fewer decimals. It would be useful for you to remember the numbers 2.58, 1.96, and 1.645.

An interesting and useful fact about the normal distribution is that a weighted sum of normal random variables has a normal distribution. That is, if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ then

$$Y = a_1X_1 + a_2X_2 \sim N(\mu_Y = a_1\mu_1 + a_2\mu_2, \sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\sigma_{12}) \tag{P.40}$$

where $\sigma_{12} = \text{cov}(X_1, X_2)$. A number of important probability distributions are related to the normal distribution. The *t*-distribution, the chi-square distribution, and the *F*-distribution are discussed in Appendix B.

TABLE P.7 Standard Normal Percentiles

α	$z_\alpha = \Phi^{-1}(\alpha)$	Rounded
0.995	2.57583	2.58
0.990	2.32635	2.33
0.975	1.95996	1.96
0.950	1.64485	1.645
0.900	1.28155	1.28
0.100	−1.28155	−1.28
0.050	−1.64485	−1.645
0.025	−1.95996	−1.96
0.010	−2.32635	−2.33
0.005	−2.57583	−2.58

P.7.1 The Bivariate Normal Distribution

Two continuous random variables, X and Y, have a **joint normal**, or **bivariate normal**, distribution if their joint *pdf* takes the form

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right] \Big/ 2(1-\rho^2) \right\}$$

where $-\infty < x < \infty$, $-\infty < y < \infty$. The parameters μ_X and μ_Y are the means of X and Y, σ_X^2 and σ_Y^2 are the variances of X and Y, so that σ_X and σ_Y are the standard deviations. The parameter ρ is the correlation between X and Y. If $\text{cov}(X, Y) = \sigma_{XY}$ then

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

The complex equation for $f(x, y)$ defines a surface in three-dimensional space. In Figure P.6a[13] we depict the surface if $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\rho = 0.7$. The positive correlation means there is a positive linear association between the values of X and Y, as described in Figure P.4. Figure P.6b depicts the contours of the density, the result of slicing the density horizontally, at a given height. The contours are more "cigar-shaped" the larger the absolute value of the correlation ρ. In Figure P.7a the correlation is $\rho = 0$. In this case the joint density is symmetrical and the contours in Figure P.7b are circles. If X and Y are jointly normal then they are statistically independent if, and only if, $\rho = 0$.

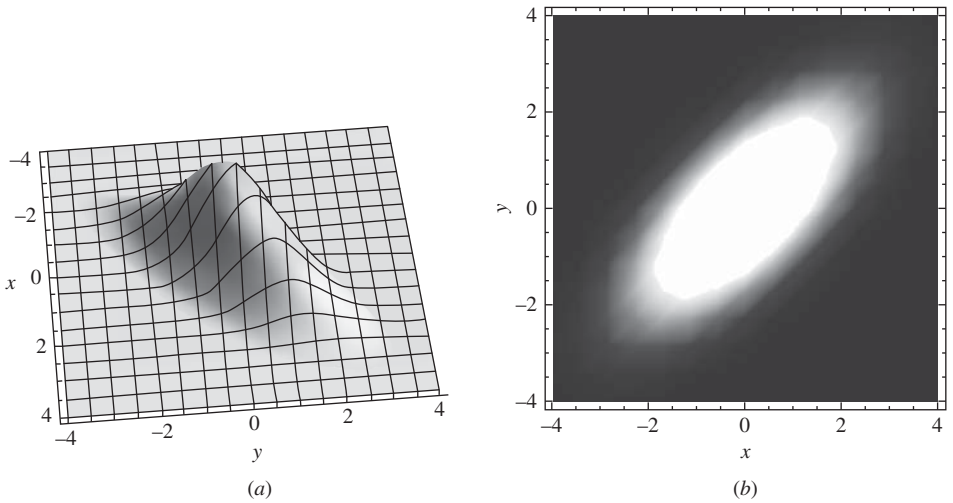


FIGURE P.6 **The bivariate normal distribution: $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\rho = 0.7$.**

[13] "The Bivariate Normal Distribution" from the Wolfram Demonstrations Project http://demonstrations.wolfram.com/. Figures P.6, P.7, and P.8 represent the interactive graphics on the site as static graphics for the primer. The site permits easy manipulation of distribution parameters. The joint density function figure can be rotated and viewed from different angles.

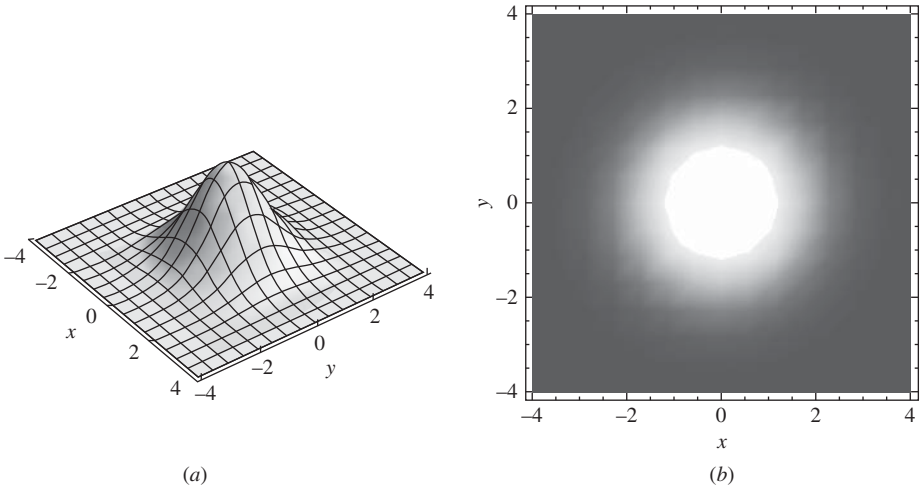


FIGURE P.7 **The bivariate normal distribution: $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\rho = 0$.**

There are several relations between the normal, bivariate normal, and the conditional distributions that are used in statistics and econometrics. First, if X and Y have a bivariate normal distribution then the marginal distributions of X and Y are normal distributions too, $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$.

Second, the conditional distribution for Y given X is normal, with conditional mean $E(Y|X) = \alpha + \beta X$, where $\alpha = \mu_Y - \beta\mu_X$ and $\beta = \sigma_{XY}/\sigma_X^2$, and conditional variance $\text{var}(Y|X) = \sigma_Y^2\left(1 - \rho^2\right)$. Or $Y|X \sim N\left[\alpha + \beta X, \sigma_Y^2\left(1 - \rho^2\right)\right]$. Three noteworthy points about these results are (i) that the conditional mean is a linear function of X, and is called a **linear regression function**; (ii) the conditional variance is constant and does not vary with X; and (iii) the conditional variance is smaller than the unconditional variance if $\rho \neq 0$. In Figure P.8[14] we display a joint normal density with

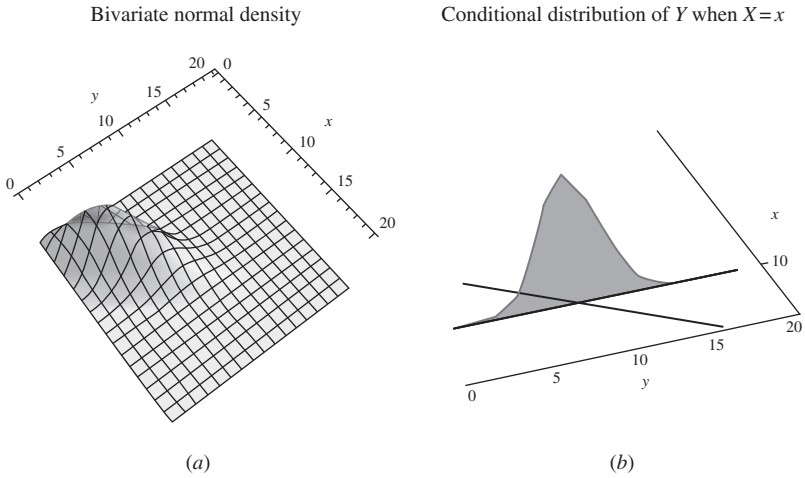


FIGURE P.8 **(*a*) Bivariate normal distribution with $\mu_X = \mu_Y = 5$, $\sigma_X = \sigma_Y = 3$, and $\rho = 0.7$; (*b*) conditional distribution of Y given $X = 10$.**

[14] "The Bivariate Normal and Conditional Distributions" from the Wolfram Demonstrations Project http://demonstrations.wolfram.com/TheBivariateNormalAndConditionalDistributions/. Both the bivariate distribution and conditional distributions can be rotated and viewed from different perspectives.

$\mu_X = \mu_Y = 5$, $\sigma_X = \sigma_Y = 3$, and $\rho = 0.7$. The covariance between X and Y is $\sigma_{XY} = \rho\sigma_X\sigma_Y = 0.7 \times 3 \times 3 = 6.3$ so that $\beta = \sigma_{XY}/\sigma_X^2 = 6.3/9 = 0.7$ and $\alpha = \mu_Y - \beta\mu_X = 5 - 0.7 \times 5 = 1.5$. The conditional mean of Y given $X = 10$ is $E(Y|X = 10) = \alpha + \beta X = 1.5 + 0.7X = 1.5 + 0.7 \times 10 = 8.5$. The conditional variance is $\text{var}(Y|X = 10) = \sigma_Y^2(1 - \rho^2) = 3^2(1 - 0.7^2) = 9(0.51) = 4.59$. That is, the conditional distribution is $(Y|X = 10) \sim N(8.5, 4.59)$.

P.8 Exercises

Answers to odd-numbered exercises are on the book website www.principlesofeconometrics.com/poe5.

P.1 Let $x_1 = 17$, $x_2 = 1$, $x_3 = 0$; $y_1 = 5$, $y_2 = 2$, $y_3 = 8$. Calculate the following:

a. $\sum_{i=1}^{2} x_i$

b. $\sum_{t=1}^{3} x_t y_t$

c. $\bar{x} = \left(\sum_{i=1}^{3} x_i\right)/3$ [Note: $\bar{x}$ is called the arithmetic average or arithmetic mean.]

d. $\sum_{i=1}^{3} (x_i - \bar{x})$

e. $\sum_{i=1}^{3} (x_i - \bar{x})^2$

f. $\left(\sum_{i=1}^{3} x_i^2\right) - 3\bar{x}^2$

g. $\sum_{i=1}^{3} (x_i - \bar{x})(y_i - \bar{y})$ where $\bar{y} = \left(\sum_{i=1}^{3} y_i\right)/3$

h. $\left(\sum_{j=1}^{3} x_j y_j\right) - 3\bar{x}\bar{y}$

P.2 Express each of the following sums in summation notation.

a. $(x_1/y_1) + (x_2/y_2) + (x_3/y_3) + (x_4/y_4)$

b. $y_2 + y_3 + y_4$

c. $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4$

d. $x_3y_5 + x_4y_6 + x_5y_7$

e. $(x_3/y_3^2) + (x_4/y_4^2)$

f. $(x_1 - y_1) + (x_2 - y_2) + (x_3 - y_3) + (x_4 - y_4)$

P.3 Write out each of the following sums and compute where possible.

a. $\sum_{i=1}^{3} (a - bx_i)$

b. $\sum_{t=1}^{4} t^2$

c. $\sum_{x=0}^{2} (2x^2 + 3x + 1)$

d. $\sum_{x=2}^{4} f(x + 3)$

e. $\sum_{x=1}^{3} f(x, y)$

f. $\sum_{x=3}^{4} \sum_{y=1}^{2} (x + 2y)$

P.4 Show algebraically that

a. $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$

b. $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}$

c. $\sum_{j=1}^{n} (x_j - \bar{x}) = 0$

P.5 Let *SALES* denote the monthly sales at a bookstore. Assume *SALES* are normally distributed with a mean of \$50,000 and a standard deviation of \$6000.

a. Compute the probability that the firm has a month with *SALES* greater than \$60,000. Show a sketch.

b. Compute the probability that the firm has a month with *SALES* between \$40,000 and \$55,000. Show a sketch.

c. Find the value of *SALES* that represents the 97th percentile of the distribution. That is, find the value $SALES_{0.97}$ such that $P(SALES > SALES_{0.97}) = 0.03$.

d. The bookstore knows their *PROFITS* are 30% of *SALES* minus fixed costs of \$12,000. Find the probability of having a month in which *PROFITS* were zero or negative. Show a sketch. [*Hint*: What is the distribution of *PROFITS*?]

P.6 A venture capital company feels that the rate of return (X) on a proposed investment is approximately normally distributed with a mean of 40% and a standard deviation of 10%.

a. Find the probability that the return X will exceed 55%.

b. The banking firm who will fund the venture sees the rate of return differently, claiming that venture capitalists are always too optimistic. They perceive that the distribution of returns is $V = 0.8X - 5\%$, where X is the rate of return expected by the venture capital company. If this is correct, find the probability that the return V will exceed 55%.

P.7 At supermarkets sales of "Chicken of the Sea" canned tuna vary from week to week. Marketing researchers have determined that there is a relationship between sales of canned tuna and the price of canned tuna. Specifically, $SALES = 50000 - 100\ PRICE$. $SALES$ is measured as the number of cans per week and $PRICE$ is measured in cents per can. Suppose $PRICE$ over the year can be considered (approximately) a normal random variable with mean $\mu = 248$ cents and standard deviation $\sigma = 10$ cents.

a. Find the expected value of $SALES$.

b. Find the variance of $SALES$.

c. Find the probability that more than 24,000 cans are sold in a week. Draw a sketch illustrating the calculation.

d. Find the $PRICE$ such that $SALES$ is at its 95th percentile value. That is, let $SALES_{0.95}$ be the 95th percentile of $SALES$. Find the value $PRICE_{0.95}$ such that $P\ (SALES > SALES_{0.95}) = 0.05$.

P.8 The Shoulder and Knee Clinic knows that their expected monthly revenue from patients depends on their level of advertising. They hire an econometric consultant who reports that their expected monthly revenue, measured in \$1000 units, is given by the following equation $E(REVENUE|ADVERT) = 100 + 20\ ADVERT$, where $ADVERT$ is advertising expenditure in \$1000 units. The econometric consultant also claims that $REVENUE$ is normally distributed with variance $\text{var}(REVENUE|ADVERT) = 900$.

a. Draw a sketch of the relationship between expected $REVENUE$ and $ADVERT$ as $ADVERT$ varies from 0 to 5.

b. Compute the probability that $REVENUE$ is greater than 110 if $ADVERT = 2$. Draw a sketch to illustrate your calculation.

c. Compute the probability that $REVENUE$ is greater than 110 if $ADVERT = 3$.

d. Find the 2.5 and 97.5 percentiles of the distribution of $REVENUE$ when $ADVERT = 2$. What is the probability that $REVENUE$ will fall in this range if $ADVERT = 2$?

e. Compute the level of $ADVERT$ required to ensure that the probability of $REVENUE$ being larger than 110 is 0.95.

P.9 Consider the U.S. population of registered voters, who may be Democrats, Republicans or independents. When surveyed about the war with ISIS, they were asked if they strongly supported war efforts, strongly opposed the war, or were neutral. Suppose that the proportion of voters in each category is given in Table P.8:

TABLE P.8 **Table for Exercise P.9**

		War Attitude		
		Against	Neutral	In Favor
Political Party	Republican	0.05	0.15	0.25
	Independent	0.05	0.05	0.05
	Democrat	0.35	0.05	0

a. Find the "marginal" probability distributions for war attitudes and political party affiliation.

b. What is the probability that a randomly selected person is a political independent given that they are in favor of the war?

c. Are the attitudes about war with ISIS and political party affiliation statistically independent or not? Why?

d. For the attitudes about the war assign the numerical values $AGAINST = 1$, $NEUTRAL = 2$, and $IN\ FAVOR = 3$. Call this variable WAR. Find the expected value and variance of WAR.

e. The Republican party has determined that monthly fundraising depends on the value of WAR from month to month. In particular the monthly contributions to the party are given by the relation (in millions of dollars) $CONTRIBUTIONS = 10 + 2 \times WAR$. Find the mean and standard deviation of $CONTRIBUTIONS$ using the rules of expectations and variance.

P.10 A firm wants to bid on a contract worth \$80,000. If it spends \$5000 on the proposal it has a 50–50 chance of getting the contract. If it spends \$10,000 on the proposal it has a 60% chance of winning the contract. Let X denote the net revenue from the contract when the \$5000 proposal is used and let Y denote the net revenue from the contract when the \$10,000 proposal is used.

X	$f(x)$
−5,000	0.5
75,000	0.5

y	$f(y)$
−10,000	0.4
70,000	0.6

a. If the firm bases its choice solely on expected value, how much should it spend on the proposal?

b. Compute the variance of X. [*Hint*: Using scientific notation simplifies calculations.]

c. Compute the variance of Y.

d. How might the variance of the net revenue affect which proposal the firm chooses?

P.11 Prior to presidential elections citizens of voting age are surveyed. In the population, two characteristics of voters are their registered party affiliation (republican, democrat, or independent) and for whom they voted in the previous presidential election (republican or democrat). Let us draw a citizen at random, defining these two variables.

$$PARTY = \begin{cases} -1 & \text{registered republican} \\ 0 & \text{independent or unregistered} \\ 1 & \text{registered democrat} \end{cases}$$

$$VOTE = \begin{cases} -1 & \text{voted republican in previous election} \\ 1 & \text{voted democratic in previous election} \end{cases}$$

a. Suppose that the probability of drawing a person who voted republication in the last election is 0.466, and the probability of drawing a person who is registered republican is 0.32, and the probability that a randomly selected person votes republican given that they are a registered republican is 0.97. Compute the joint probability Prob[$PARTY = -1$, $VOTE = -1$]. Show your work.

b. Are these random variables statistically independent? Explain.

P.12 Based on years of experience, an economics professor knows that on the first principles of economics exam of the semester 13% of students will receive an A, 22% will receive a B, 35% will receive a C, 20% will receive a D, and the remainder will earn an F. Assume a 4 point grading scale (A = 4, B = 3, C = 2, D = 1, and F = 0). Define the random variable $GRADE = 4, 3, 2, 1, 0$ to be the grade of a randomly chosen student.

a. What is the probability distribution $f(GRADE)$ for this random variable?

b. What is the expected value of $GRADE$? What is the variance of $GRADE$? Show your work.

c. The professor has 300 students in each class. Suppose that the grade of the ith student is $GRADE_i$ and that the probability distribution of grades $f(GRADE_i)$ is the same for all students. Define $CLASS\_AVG = \sum_{i=1}^{300} GRADE_i/300$. Find the expected value and variance of $CLASS\_AVG$.

d. The professor has estimated that the number of economics majors coming from the class is related to the grade on the first exam. He believes the relationship to be $MAJORS = 50 + 10CLASS\_AVG$. Find the expected value and variance of $MAJORS$. Show your work.

P.13 The LSU Tigers baseball team will play the Alabama baseball team in a weekend series of two games. Let $W = 0$, 1, or 2 equal the number of games LSU wins. Let the weekend's weather be designated as Cold or Not Cold. Let $C = 1$ if the weather is cold and $C = 0$ if the weather is not cold. The joint probability function of these two random variables is given in Table P.9, along with space for the marginal distributions.

TABLE P.9 **Table for Exercise P.13**

	$W = 0$	$W = 1$	$W = 2$	$f(c)$
$C = 1$	(i)	0.12	0.12	(ii)
$C = 0$	0.07	0.14	(iii)	(iv)
$f(w)$	(v)	(vi)	0.61	

a. Fill in the blanks, (i)–(vi).

b. Using the results of (a), find the conditional probability distribution of the number of wins, W, conditional on the weather being warm, $C = 0$. Based on a comparison of the conditional probability distribution $f(w|C = 0)$ and the marginal distribution $f(w)$, can you conclude that the number of games LSU wins W is statistically independent of the weather conditions C, or not? Explain.

c. Find the expected value of the number of LSU wins, W. Also find the conditional expectation $E(W|C = 0)$. Show your work. What kind of weather is more favorable for the LSU Tigers baseball team?

d. The revenue of vendors at the LSU Alex Box baseball stadium depends on the crowds, which in turn depends on the weather. Suppose that food sales $FOOD = \$10{,}000 - 3000C$. Use the rules for expected value and variance to find the expected value and standard deviation of food sales.

P.14 A clinic specializes in shoulder injuries. A patient is randomly selected from the population of all clinic clients. Let S be the number of doctor visits for shoulder problems in the past six months. Assume the values of S are $s = 1, 2, 3,$ or 4. Patients at the shoulder clinic are also asked about knee injuries. Let K = the number of doctor visits for knee injuries during the past six months. Assume the values of K are $k = 0$, 1 or 2. The joint probability distribution of the numbers of shoulder and knee injuries is shown in Table P.10. Use the information in the joint probability distribution to answer the following questions. Show **brief** calculations for each

TABLE P.10 **Table for Exercise P.14**

		Knee = K			
		0	1	2	$f(s)$
Shoulder = S	1	0.15	0.09	0.06	
	2	0.06			
	3	0.02	0.10		0.2
	4	0.02	0.08	0.10	
	$f(k)$			0.33	

a. What is the probability that a randomly chosen patient will have two doctor visits for shoulder problems during the past six months?

b. What is the probability that a randomly chosen patient will have two doctor visits for shoulder problems during the past six months given that they have had one doctor visit for a knee injury in the past six months?

c. What is the probability that a randomly chosen patient will have had three doctor visits for shoulder problems and two doctor visits for knee problems in the past six months?

d. Are the number of doctor visits for knee and shoulder injuries statistically independent? Explain.

e. What is the expected value of the number of doctor visits for shoulder injuries from this population?

f. What is the variance of the number of doctor visits for shoulder injuries from this population?

P.15 As you walk into your econometrics exam, a friend bets you \$20 that she will outscore you on the exam. Let X be a random variable denoting your winnings. X can take the values 20, 0 [if there is a tie], or -20. You know that the probability distribution for X, $f(x)$, depends on whether she studied for

the exam or not. Let $Y = 0$ if she studied and $Y = 1$ if she did not study. Consider the following joint distribution Table P.11.

TABLE P.11 **Joint *pdf* for Exercise P.15**

		Y		
		0	**1**	$f(x)$
X	−20	(i)	0	(ii)
	0	(iii)	0.15	0.25
	20	0.10	(iv)	(v)
	$f(y)$	(vi)	0.60	

a. Fill in the missing elements (i)–(vi) in the table.
b. Compute $E(X)$. Should you take the bet?
c. What is the probability distribution of your winnings if you know that she did not study?
d. Find your expected winnings given that she did not study.
e. Use the Law of Iterated Expectations to find $E(X)$.

P.16 Breast cancer prevalence in the United Kingdom can be summarized for the population (data are in 1000s) as in Table P.12.

TABLE P.12 **Table for Exercise P.16**

	Sex		
	Female	**Male**	**Total**
Suffers from Breast Cancer	550	3	553
Not Suffering from Breast Cancer	30,868	30,371	61,239
Total	31,418	30,374	61,792

a. Compute the probability that a randomly drawn person has breast cancer.
b. Compute the probability that a randomly drawn female has breast cancer.
c. Compute the probability that a person is female given that the person has breast cancer.
d. What is the conditional probability function for the prevalence of breast cancer given that the person is female?
e. What is the conditional probability function for the prevalence of breast cancer given that the person is male?

P.17 A continuous random variable Y has *pdf*

$$f(y) = \begin{cases} 2y & 0 < y < 1 \\ 0 & \textit{otherwise} \end{cases}$$

a. Sketch the *pdf*.
b. Find the *cdf*, $F(y) = P(Y \leq y)$ and sketch it. [*Hint*: Requires calculus.]
c. Use the *pdf* and a geometric argument to find the probability $P(Y \leq 1/2)$.
d. Use the *cdf* from part (b) to compute $P(Y \leq 1/2)$.
e. Using the *pdf* and a geometric argument find the probability $P(1/4 \leq Y \leq 3/4)$.
f. Use the *cdf* from part (b) to compute $P(1/4 \leq Y \leq 3/4)$.

P.18 Answer each of the following:

a. An internal revenue service auditor knows that 3% of all income tax forms contain errors. Returns are assigned randomly to auditors for review. What is the probability that an auditor will have to

view four tax returns until the first error is observed? That is, what is the probability of observing three returns with no errors, and then observing an error in the fourth return?

b. Let Y be the number of independent trials of an experiment before a success is observed. That is, it is the number of failures before the first success. Assume each trial has a probability of success of p and a probability of failure of $1 - p$. Is this a discrete or continuous random variable? What is the set of possible values that Y can take? Can Y take the value zero? Can Y take the value 500?

c. Consider the *pdf* $f(y) = P(Y = y) = p(1 - p)^y$. Using this *pdf* compute the probability in (a). Argue that this probability function generally holds for the experiment described in (b).

d. Using the value $p = 0.5$, plot the *pdf* in (c) for $y = 0, 1, 2, 3, 4$.

e. Show that $\sum_{y=0}^{\infty} f(y) = \sum_{y=0}^{\infty} p(1 - p)^y = 1$. [*Hint*: If $|r| < 1$ then $1 + r + r^2 + r^3 + \cdots = 1/(1-r)$.]

f. Verify for $y = 0, 1, 2, 3, 4$ that the *cdf* $P(Y \le y) = 1 - (1 - p)^{y+1}$ yields the correct values.

P.19 Let X and Y be random variables with expected values $\mu = \mu_X = \mu_Y$ and variances $\sigma^2 = \sigma_X^2 = \sigma_Y^2$. Let $Z = (2X + Y)/2$.

a. Find the expected value of Z.

b. Find the variance of Z assuming X and Y are statistically independent.

c. Find the variance of Z assuming that the correlation between X and Y is -0.5.

d. Let the correlation between X and Y be -0.5. Find the correlation between aX and bY, where a and b are any nonzero constants.

P.20 Suppose the *pdf* of the continuous random variable X is $f(x) = 1$, for $0 < x < 1$ and $f(x) = 0$ otherwise.

a. Draw a sketch of the *pdf*. Verify that the area under the *pdf* for $0 < x < 1$ is 1.

b. Find the *cdf* of X. [*Hint*: Requires the use of calculus.]

c. Compute the probability that X falls in each of the intervals [0, 0.1], [0.5, 0.6], and [0.79, 0.89]. Indicate the probabilities on the sketch drawn in (a).

d. Find the expected value of X.

e. Show that the variance of X is 1/12.

f. Let Y be a discrete random variable taking the values 1 and 0 with conditional probabilities $P(Y = 1|X = x) = x$ and $P(Y = 0|X = x) = 1 - x$. Use the Law of Iterated Expectations to find $E(Y)$.

g. Use the variance decomposition to find $\text{var}(Y)$.

P.21 A fair die is rolled. Let Y be the face value showing, 1, 2, 3, 4, 5, or 6 with each having the probability 1/6 of occurring. Let X be another random variable that is given by

$$X = \begin{cases} Y & \text{if } Y \text{ is even} \\ 0 & \text{if } Y \text{ is odd} \end{cases}$$

a. Find $E(Y)$, $E\left(Y^2\right)$, and $\text{var}(Y)$.

b. What is the probability distribution for X? Find $E(X)$, $E\left(X^2\right)$, and $\text{var}(X)$.

c. Find the conditional probability distribution of Y given each X.

d. Find the conditional expected value of Y given each value of X, $E(Y|X)$.

e. Find the probability distribution of $Z = XY$. Show that $E(Z) = E(XY) = E\left(X^2\right)$.

f. Find $\text{cov}(X, Y)$.

P.22 A large survey of married women asked "How many extramarital affairs did you have last year?" 77% said they had none, 5% said they had one, 2% said two, 3% said three, and the rest said more than three. Assume these women are representative of the entire population.

a. What is the probability that a randomly selected married woman will have had one affair in the past year?

b. What is the probability that a randomly selected married woman will have had more than one affair in the past year?

c. What is the probability that a randomly chosen married woman will have had less than three affairs in the past year?

d. What is the probability that a randomly chosen married woman will have had one or two affairs in the past year?

e. What is the probability that a randomly chosen married woman will have had one or two affairs in the past year, given that they had at least one?

P.23 Let *NKIDS* represent the number of children ever born to a woman. The possible values of *NKIDS* are $nkids = 0, 1, 2, 3, 4, \ldots$. Suppose the *pdf* is $f(nkids) = 2^{nkids}/(7.389nkids!)$, where ! denotes the factorial operation.

- **a.** Is *NKIDS* a discrete or continuous random variable?
- **b.** Calculate the *pdf* for $nkids = 0, 1, 2, 3, 4$. Sketch it. [Note: It may be convenient to use a spreadsheet or other software to carry out tedious calculations.]
- **c.** Calculate the probabilities $P[NKIDS \leq nkids]$ for $nkids = 0, 1, 2, 3, 4$. Sketch the cumulative distribuiton function.
- **d.** What is the probability that a woman will have more than one child.
- **e.** What is the probability that a woman will have two or fewer children?

P.24 Five baseballs are thrown to a batter who attempts to hit the ball 350 feet or more. Let H denote the number of successes, with the *pd* for having h successes being $f(h) = 120 \times 0.4^h \times 0.6^{5-h}/[h!(5-h)!]$, where ! denotes the factorial operation.

- **a.** Is H a discrete or continuous random variable? What values can it take?
- **b.** Calculate the probabilities that the number of successes $h = 0, 1, 2, 3, 4$, and 5. [Note: It may be convenient to use a spreadsheet or other software to carry out tedious calculations.] Sketch the *pdf*.
- **c.** What is the probability of two or fewer successes?
- **d.** Find the expected value of the random variable H. Show your work.
- **e.** The prizes are \$1000 for the first success, \$2000 for the second success, \$3000 for the third success, and so on. What is the *pdf* for the random variable *PRIZE*, which is the total prize winnings?
- **f.** Find the expected value of total prize winnings, *PRIZE*.

P.25 An author knows that a certain number of typographical errors (0, 1, 2, 3, …) are on each book page. Define the random variable T equaling the number of errors per page. Suppose that T has a Poisson distribution [Appendix B.3.3], with *pdf*, $f(t) = \mu^t \exp(-\mu)/t!$, where ! denotes the factorial operation, and $\mu = E(T)$ is the mean number of typographical errors per page.

- **a.** If $\mu = 3$, what is the probability that a page has one error? What is the probability that a page has four errors?
- **b.** An editor independently checks each word of every page and catches 90% of the errors, but misses 10%. Let Y denote the number of errors caught on a page. The values of y must be less than or equal to the actual number t of errors on the page. Suppose that the number of errors caught on a page with t errors has a binomial distribution [Appendix B.3.2].

$$g(y|t,\, p = 0.9) = \frac{t!}{y!(t-y)!} 0.9^y 0.1^{t-y},\ y = 0, 1, \ldots, t$$

  Compute the probability that the editor finds one error on a page given that the page actually has four errors.
- **c.** Find the **joint** probability $P[Y = 3, T = 4]$.
- **d.** It can be shown that the probability the editor will find Y errors on a page follows a Poisson distribution with mean $E(Y) = 0.9\mu$. Use this information to find the conditional probability that there are $T = 4$ errors on a page given that $Y = 3$ are found.

CHAPTER 2

The Simple Linear Regression Model

LEARNING OBJECTIVES

Remark

Learning Objectives and *Keywords* sections will appear at the beginning of each chapter. We urge you to think about and possibly write out answers to the questions, and make sure you recognize and can define the keywords. If you are unsure about the questions or answers, consult your instructor. When examples are requested in *Learning Objectives* sections, you should think of examples *not* in the book.

Based on the material in this chapter you should be able to

1. Explain the difference between an estimator and an estimate, and why the least squares estimators are random variables, and why least squares estimates are not.
2. Discuss the interpretation of the slope and intercept parameters of the simple regression model, and sketch the graph of an estimated equation.
3. Explain the theoretical decomposition of an observable variable y into its systematic and random components, and show this decomposition graphically.
4. Discuss and explain each of the assumptions of the simple linear regression model.
5. Explain how the least squares principle is used to fit a line through a scatter plot of data. Be able to define the least squares residual and the least squares fitted value of the dependent variable and show them on a graph.
6. Define the elasticity of y with respect to x and explain its computation in the simple linear regression model when y and x are not transformed in any way, and when y and/or x have been transformed to model a nonlinear relationship.
7. Explain the meaning of the statement ''If regression model assumptions SR1–SR5 hold, then the least squares estimator b_2 is unbiased.'' In particular, what exactly does ''unbiased'' mean? Why is b_2 biased if an important variable has been omitted from the model?
8. Explain the meaning of the phrase ''sampling variability.''
9. Explain how the factors σ^2, $\sum(x_i - \bar{x})^2$, and N affect the precision with which we can estimate the unknown parameter β_2.

10. State and explain the Gauss–Markov theorem.
11. Use the least squares estimator to estimate nonlinear relationships and interpret the results.
12. Explain the difference between an explanatory variable that is fixed in repeated samples and an explanatory variable that is random.
13. Explain the term ''random sampling.''

KEYWORDS

assumptions
asymptotic
biased estimator
BLUE
degrees of freedom
dependent variable
deviation from the mean form
econometric model
economic model
elasticity
exogenous variable
Gauss–Markov theorem
heteroskedastic
homoskedastic
independent variable
indicator variable
least squares estimates
least squares estimators
least squares principle
linear estimator
log-linear model
nonlinear relationship
prediction
quadratic model
random error term
random-*x*
regression model
regression parameters
repeated sampling
sampling precision
sampling properties
scatter diagram
simple linear regression analysis
simple linear regression function
specification error
strictly exogenous
unbiased estimator

Economic theory suggests many relationships between economic variables. In microeconomics, you considered demand and supply models in which the quantities demanded and supplied of a good depend on its price. You considered "production functions" and "total product curves" that explained the amount of a good produced as a function of the amount of an input, such as labor, that is used. In macroeconomics, you specified "investment functions" to explain that the amount of aggregate investment in the economy depends on the interest rate and "consumption functions" that related aggregate consumption to the level of disposable income.

Each of these models involves a relationship between economic variables. In this chapter, we consider how to use a sample of economic data to quantify such relationships. As economists, we are interested in questions such as the following: If one variable (e.g., the price of a good) changes in a certain way, *by how much* will another variable (the quantity demanded or supplied) change? Also, given that we know the value of one variable, can we *forecast* or *predict* the corresponding value of another? We will answer these questions by using a **regression model**. Like all models, the regression model is based on **assumptions**. In this chapter, we hope to be very clear about these assumptions, as they are the conditions under which the analysis in subsequent chapters is appropriate.

2.1 An Economic Model

In order to develop the ideas of regression models, we are going to use a simple, but important, economic example. Suppose that we are interested in studying the relationship between household income and expenditure on food. Consider the "experiment" of randomly selecting households from a particular population. The population might consist of households within a particular city, state, province, or country. For the present, suppose that we are interested only in households with an income of $1000 per week. In this experiment, we randomly select a number of households from this population and interview them. We ask the question "How much did you spend per person on food last week?" Weekly food expenditure, which we denote as y, is a *random variable* since the value is unknown to us until a household is selected and the question is asked and answered.

Remark

In the Probability Primer and Appendices B and C, we distinguished random variables from their values by using uppercase (Y) letters for random variables and lowercase (y) letters for their values. We *will not* make this distinction any longer because it leads to complicated notation. We will use lowercase letters, like "y," to denote random variables as well as their values, and we will make the interpretation clear in the surrounding text.

The continuous random variable y has a probability density function (which we will abbreviate as *pdf*) that describes the probabilities of obtaining various food expenditure values. *If you are rusty or uncertain about probability concepts, see the Probability Primer and Appendix B at the end of this book for a comprehensive review*. The amount spent on food per person will vary from one household to another for a variety of reasons: some households will be devoted to gourmet food, some will contain teenagers, some will contain senior citizens, some will be vegetarian, and some will eat at restaurants more frequently. All of these factors and many others, including random, impulsive buying, will cause weekly expenditures on food to vary from one household to another, even if they all have the same income. The *pdf* $f(y)$ describes how expenditures are "distributed" over the population and might look like Figure 2.1.

The *pdf* in Figure 2.1a is actually a conditional *pdf* since it is "conditional" upon household income. If x = weekly household income = \$1000, then the conditional *pdf* is $f(y|x = \$1000)$. The *conditional mean*, or *expected value*, of y is $E(y|x = \$1000) = \mu_{y|x}$ and is our population's mean weekly food expenditure per person.

Remark

The expected value of a random variable is called its "mean" value, which is really a contraction of *population mean*, the center of the probability distribution of the random variable. This is *not* the same as the *sample mean*, which is the arithmetic average of numerical values. Keep the distinction between these two usages of the term "mean" in mind.

The *conditional variance* of y is $\text{var}(y|x = \$1000) = \sigma^2$, which measures the dispersion of household expenditures y about their mean $\mu_{y|x}$. The parameters $\mu_{y|x}$ and σ^2, if they were known, would give us some valuable information about the population we are considering. If we knew these parameters, and if we knew that the conditional distribution $f(y|x = \$1000)$ was *normal*,

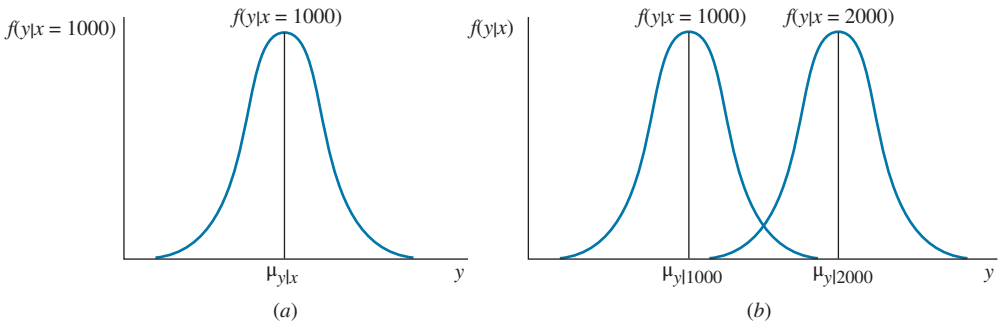


FIGURE 2.1 **(*a*) Probability distribution $f(y|x = 1000)$ of food expenditure y given income $x = \$1000$. (*b*) Probability distributions of food expenditure y given incomes $x = \$1000$ and $x = \$2000$.**

$N(\mu_{y|x}, \sigma^2)$, then we could calculate probabilities that y falls in specific intervals using properties of the normal distribution. That is, we could compute the proportion of the household population that spends between \$50 and \$75 per person on food, given \$1000 per week income.

As economists, we are usually interested in studying relationships between variables, in this case the relationship between y = weekly food expenditure per person and x = weekly household income. Economic theory tells us that expenditure on economic goods depends on income. Consequently, we call y the "**dependent variable**" and x the "**independent**" or "**explanatory**" variable. In econometrics, we recognize that real-world expenditures are random variables, and we want to use data to learn about the relationship.

An econometric analysis of the expenditure relationship can provide answers to some important questions, such as: If weekly income goes up by \$100, **how much** will average weekly food expenditures rise? Or, could weekly food expenditures fall as income rises? How much would we predict the weekly per person expenditure on food to be for a household with an income of \$2000 per week? The answers to such questions provide valuable information for decision makers.

> *Using … per person food spending information … one can determine the similarities and disparities in the spending habits of households of differing sizes, races, incomes, geographic areas, and other socioeconomic and demographic features. This information is valuable for assessing existing market conditions, product distribution patterns, consumer buying habits, and consumer living conditions. Combined with demographic and income projections, this information may be used to anticipate consumption trends. The information may also be used to develop typical market baskets of food for special population groups, such as the elderly. These market baskets may, in turn, be used to develop price indices tailored to the consumption patterns of these population groups*. [*Blisard, Noel, Food Spending in American Households, 1997–1998, Electronic Report from the Economic Research Service, U.S. Department of Agriculture, Statistical Bulletin Number 972, June 2001*]

From a business perspective, if we are managers of a supermarket chain (or restaurant, or health food store, etc.), we must consider long-range plans. If economic forecasters are predicting that local income will increase over the next few years, then we must decide whether, and how much, to expand our facilities to serve our customers. Or, if we plan to open franchises in high-income and low-income neighborhoods, then forecasts of expenditures on food per person, along with neighborhood demographic information, give an indication of how large the stores in those areas should be.

In order to investigate the relationship between expenditure and income, we must build an **economic model** and then a corresponding **econometric model** that forms the basis for a quantitative or *empirical* economic analysis. In our food expenditure example, economic theory suggests that average weekly per person household expenditure on food, represented mathematically by the conditional mean $E(y|x) = \mu_{y|x}$, depends on household income x. If we consider households with different levels of income, we expect the average expenditure on food to change. In Figure 2.1b, we show the *pdfs* of food expenditure for two different levels of weekly income, \$1000 and \$2000. Each conditional *pdf* $f(y|x)$ shows that expenditures will be distributed about a mean value $\mu_{y|x}$, but the mean expenditure by households with higher income is larger than the mean expenditure by lower income households.

In order to use data, we must now specify an *econometric model* that describes how the data on household income and food expenditure are obtained and that guides the econometric analysis.

2.2 An Econometric Model

Given the economic reasoning in the previous section, and to quantify the relationship between food expenditure and income, we must progress from the ideas in Figure 2.1, to an **econometric model**. First, suppose a three-person household has an unwavering rule that each week they

spend \$80 and then also spend 10 cents of each dollar of income received on food. Let y = weekly household food expenditure (\$) and let x = weekly household income (\$). Algebraically their rule is $y = 80 + 0.10x$. Knowing this relationship we calculate that in a week in which the household income is \$1000, the household will spend \$180 on food. If weekly income increases by \$100 to \$1100, then food expenditure increases to \$190. These are **predictions** of food expenditure given income. **Predicting** the value of one variable given the value of another, or others, is one of the primary uses of regression analysis.

A second primary use of regression analysis is to attribute, or relate, changes in one variable to changes in another variable. To that end, let "Δ" denote "change in" in the usual algebraic way. A change in income of \$100 means that $\Delta x = 100$. Because of the spending rule $y = 80 + 0.10x$ the change in food expenditure is $\Delta y = 0.10\Delta x = 0.10(100) = 10$. An increase in income of \$100 leads to, or causes, a \$10 increase in food expenditure. Geometrically, the rule is a line with "y-intercept" 80 and slope $\Delta y/\Delta x = 0.10$. An economist might say that the household "marginal propensity to spend on food is 0.10," which means that from each additional dollar of income 10 cents is spent on food. Alternatively, in a kind of economist shorthand, the "marginal effect of income on food expenditure is 0.10." Much of economic and econometric analysis is an attempt to measure a **causal relationship** between two economic variables. Claiming **causality** here, that is, changing income leads to a change in food expenditure, is quite clear given the household's expenditure rule. It is not always so straightforward.

In reality, many other factors may affect household expenditure on food; the ages and sexes of the household members, their physical size, whether they do physical labor or have desk jobs, whether there is a party following the big game, whether it is an urban or rural household, whether household members are vegetarians or into a paleo-diet, as well as other taste and preference factors ("I really like truffles") and impulse shopping ("Wow those peaches look good!"). Lots of factors. Let e = *everything else* affecting food expenditure other than income. Furthermore, even if a household has a rule, strict or otherwise, we do not know it. To account for these realities, we suppose that the household's food expenditure decision is based on the equation

$$y = \beta_1 + \beta_2 x + e \tag{2.1}$$

In addition to y and x, equation (2.1) contains two unknown **parameters**, β_1 and β_2, instead of "80" and "0.10," and an **error term** e, which represents all those other factors (*everything else*) affecting weekly household food expenditure.

Imagine that we can perform an experiment on the household. Let's increase the household's income by \$100 per week and hold other things constant. Holding other things constant, or holding all else (*everything else*) equal, is the *ceteris paribus* assumption discussed extensively in economic principles courses. Let $\Delta x = 100$ denote the change in household income. Assuming everything else affecting household food expenditure, e, is held constant means that $\Delta e = 0$. The effect of the change in income is $\Delta y = \beta_2 \Delta x + \Delta e = \beta_2 \Delta x = \beta_2 \times 100$. The change in weekly food expenditure $\Delta y = \beta_2 \times 100$ is explained by, or caused by, the change in income. The unknown parameter β_2, the marginal propensity to spend on food from income, tells us the proportion of the increase in income used for food purchases; it answers the "how much" question "How much will food expenditure change given a change in income, holding all else constant?"

The experiment in the previous paragraph is not feasible. We can give a household an extra \$100 income, but we cannot hold all else constant. The simple calculation of the marginal effect of an increase in income on food expenditure $\Delta y = \beta_2 \times 100$ is not possible. However, we can shed light on this "how much" question by using **regression analysis** to estimate β_2. Regression analysis is a statistical method that uses data to explore relationships between variables. A **simple linear regression analysis** examines the relationship between a y-variable and one x-variable. It is said to be "simple" not because it is easy, but because there is only one x-variable. The y-variable is called the dependent variable, the outcome variable, the explained variable, the left-hand-side variable, or the regressand. In our example, the dependent variable is

y = weekly household expenditure on food. The variable x = weekly household income is called the independent variable, the explanatory variable, the right-hand-side variable, or the regressor. Equation (2.1) is the **simple linear regression** model.

All models are abstractions from reality and working with models requires assumptions. The same is true for the regression model. The first assumption of the simple linear regression model is that relationship (2.1) holds for the members of the population under consideration. For example, define the population to be three-person households in a given geographic region, say southern Australia. The unknowns β_1 and β_2 are called **population parameters**. We assert the behavioral rule $y = \beta_1 + \beta_2 x + e$ holds for all households in the population. Each week food expenditure equals β_1, plus a proportion β_2 of income, plus other factors, e.

The field of statistics was developed because, in general, populations are large, and it is impossible (or impossibly costly) to examine every population member. The population of three-person households in a given geographic region, even if it is only a medium-sized city, is too large to survey individually. Statistical and econometric methodology examines and analyzes a **sample of data** from the population. After analyzing the data, we make **statistical inferences**. These are conclusions or judgments about a population based on the data analysis. Great care must be taken when drawing inferences. The inferences are conclusions about the particular population from which the data were collected. Data on households from southern Australia may, or may not, be useful for making inferences, drawing conclusions, about households from the southern United States. Do Melbourne, Australia, households have the same food spending patterns as households in New Orleans, Louisiana? That might be an interesting research topic. If not, then we may not be able to draw valid conclusions about New Orleans household behavior from the sample of Australian data.

2.2.1 Data Generating Process

The sample of data, and how the data are actually obtained, is crucially important for subsequent inferences. The exact mechanisms for collecting a sample of data are very discipline specific (e.g., agronomy is different from economics) and beyond the scope of this book.[1] For the household food expenditure example, let us assume that we can obtain a sample at a point in time [these are **cross-sectional data**] consisting of N data pairs that are **randomly** selected from the population. Let (y_i, x_i) denote the ith data pair, $i = 1, \ldots, N$. The variables y_i and x_i are **random variables**, because their values are not known until they are observed. Randomly selecting households makes the first observation pair (y_1, x_1) statistically independent of all other data pairs, and each observation pair (y_i, x_i) is **statistically independent** of every other data pair, (y_j, x_j), where $i \neq j$. We further assume that the random variables y_i and x_i have a joint $pdf\, f(y_i, x_i)$ that describes their distribution of values. We often do not know the exact nature of the joint distribution (such as bivariate normal; see Probability Primer, Section P.7.1), but all pairs drawn from the same population are assumed to follow the same joint pdf, and, thus, the data pairs are not only statistically independent but are also **identically distributed** (abbreviated **i.i.d.** or *iid*). Data pairs that are *iid* are said to be a **random sample**.

If our first assumption is true, that the behavioral rule $y = \beta_1 + \beta_2 x + e$ holds for all households in the population, then restating (2.1) for each (y_i, x_i) data pair

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N \tag{2.1}$$

This is sometimes called the **data generating process** (**DGP**) because we assume that the observable data follow this relationship.

[1] See, for example, Paul S. Levy and Stanley Lemeshow (2008) *Sampling of Populations: Methods and Applications, 4th Edition*, Hoboken, NJ: John Wiley and Sons, Inc.

2.2.2 The Random Error and Strict Exogeneity

The second assumption of the simple regression model (2.1) concerns the "everything else" term e. The variables (y_i, x_i) are random variables because we do not know what values they take until a particular household is chosen and they are observed. The error term e_i is also a random variable. All the other factors affecting food expenditure except income will be different for each population household if for no other reason that everyone's tastes and preferences are different. Unlike food expenditure and income, the **random error term** e_i is **not observable**; it is **unobservable**. We cannot measure tastes and preferences in any direct way, just as we cannot directly measure the economic "utility" derived from eating a slice of cake. The second regression assumption is that the x-variable, income, cannot be used to predict the value of e_i, the effect of the collection of all other factors affecting the food expenditure by the ith household. Given an income value x_i for the ith household, the *best* (optimal) predictor[2] of the random error e_i is the conditional expectation, or conditional mean, $E(e_i|x_i)$. The assumption that x_i cannot be used to predict e_i is equivalent to saying $E(e_i|x_i) = 0$. That is, given a household's income we cannot do any better than predicting that the random error is zero; the effects of all other factors on food expenditure average out, in a very specific way, to zero. We will discuss other situations in which this might or might not be true in Section 2.10. For now, recall from the Probability Primer, Section P.6.5, that $E(e_i|x_i) = 0$ has two implications. The first is $E(e_i|x_i) = 0 \Longrightarrow E(e_i) = 0$; if the conditional expected value of the random error is zero, then the **unconditional expectation** of the random error is also zero. In the population, the average effect of all the omitted factors summarized by the random error term is zero.

The second implication is $E(e_i|x_i) = 0 \Longrightarrow \text{cov}(e_i, x_i) = 0$. If the conditional expected value of the random error is zero, then e_i, the random error for the ith observation, has covariance zero and correlation zero, with the corresponding observation x_i. In our example, the random component e_i, representing all factors affecting food expenditure except income for the ith household, is uncorrelated with income for that household. You might wonder how that could possibly be shown to be true. After all, e_i is unobservable. The answer is that it is very hard work. You must convince yourself and your audience that anything that might have been omitted from the model is not correlated with x_i. The primary tool is economic reasoning: your own intellectual experiments (i.e., thinking), reading literature on the topic and discussions with colleagues or classmates. And we really can't prove that $E(e_i|x_i) = 0$ is true with absolute certainty in most economic models.

We noted that $E(e_i|x_i) = 0$ has two implications. If either of the implications is **not** true, then $E(e_i|x_i) = 0$ is not true, that is,

$$E(e_i|x_i) \neq 0 \text{ if (i) } E(e_i) \neq 0 \text{ or if (ii) } \text{cov}(e_i, x_i) \neq 0$$

In the first case, if the population average of the random errors e_i is not zero, then $E(e_i|x_i) \neq 0$. In a certain sense, we will be able to work around the case when $E(e_i) \neq 0$, say if $E(e_i) = 3$, as you will see below. The second implication of $E(e_i|x_i) = 0$ is that $\text{cov}(e_i, x_i) = 0$; the random error for the ith observation has zero covariance and correlation with the ith observation on the explanatory variable. If $\text{cov}(e_i, x_i) = 0$, the explanatory variable x is said to be **exogenous**, providing our first assumption that the pairs (y_i, x_i) are *iid* holds. When x is exogenous, regression analysis can be used successfully to estimate β_1 and β_2. To differentiate the weaker condition $\text{cov}(e_i, x_i) = 0$, simple **exogeneity**, from the stronger condition $E(e_i|x_i) = 0$, we say that x is **strictly exogenous** if $E(e_i|x_i) = 0$. If $\text{cov}(e_i, x_i) \neq 0$, then x is said to be **endogenous**. When x is endogenous, it is more difficult, sometimes much more difficult, to carry out statistical inference. A great deal will be said about exogeneity and strict exogeneity in the remainder of this book.

[2] You will learn about optimal prediction in Appendix 4C.

EXAMPLE 2.1 | A Failure of the Exogeneity Assumption

Consider a regression model exploring the relationship between a working person's wage and their years of education, using a random sample of data. The simple regression model is $WAGE_i = \beta_1 + \beta_2 EDUC_i + e_i$, where $WAGE_i$ is the hourly wage rate of the ith randomly selected person and $EDUC_i$ is their years of education. The pairs $(WAGE_i, EDUC_i)$ from the random sample are assumed to be *iid*. In this model, the random error e_i accounts for all those factors other than $EDUC_i$ that affect a person's wage rate. What might some of those factors be? Ability, intelligence, perseverance, and industriousness are all important characteristics of an employee and likely to influence their wage rate. Are any of these factors which are bundled into e_i likely to be correlated with $EDUC_i$? A few moments reflection will lead you to say "yes." It is very plausible that those with higher education have higher ability, intelligence, perseverance, and industriousness. Thus, there is a strong argument that $EDUC_i$ is an endogenous regressor in this regression and that the strict exogeneity assumption fails.

2.2.3 The Regression Function

The importance of the strict exogeneity assumption is the following. If the strict exogeneity assumption $E(e_i|x_i) = 0$ is true, then the conditional expectation of y_i given x_i is

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i + E(e_i|x_i) = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, N \tag{2.2}$$

The conditional expectation $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ in (2.2) is called the **regression function**, or **population regression function**. It says that in the population the average value of the dependent variable for the ith observation, conditional on x_i, is given by $\beta_1 + \beta_2 x_i$. It also says that given a change in x, Δx, the resulting change in $E(y_i|x_i)$ is $\beta_2 \Delta x$ **holding all else constant**, in the sense that given x_i the average of the random errors is zero, and any change in x is not correlated with any corresponding change in the random error e. In this case, we can say that a change in x leads to, or **causes**, a change in the expected (population average) value of y given x_i, $E(y_i|x_i)$.

The regression function in (2.2) is shown in Figure 2.2, with y-intercept $\beta_1 = E(y_i|x_i = 0)$ and slope

$$\beta_2 = \frac{\Delta E(y_i|x_i)}{\Delta x_i} = \frac{dE(y_i|x_i)}{dx_i} \tag{2.3}$$

where Δ denotes "change in" and $dE(y|x)/dx$ denotes the "derivative" of $E(y|x)$ with respect to x. We will not use derivatives to any great extent in this book, and if you are not too familiar with the concept you can think of "d" as a stylized version of Δ and go on. See Appendix A.3 for a discussion of derivatives.

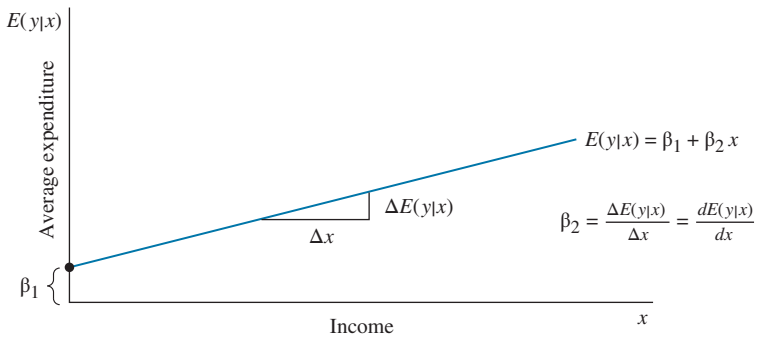


FIGURE 2.2 **The economic model: a linear relationship between average per person food expenditure and income.**

EXAMPLE 2.2 | Strict Exogeneity in the Household Food Expenditure Model

The strict exogeneity assumption is that the average of *everything else* affecting the food expenditure of the ith household, given the income of the ith household, is zero. Could this be true? One test of this possibility is the question "Using the income of the ith household, can we predict the value of e_i, the combined influence of all factors affecting food expenditure other than income?" If the answer is yes, then strict exogeneity fails. If not, then $E(e_i|x_i) = 0$ *may be* a plausible assumption. And if it is, then equation (2.1) can be interpreted as a causal model, and β_2 can be thought of as the marginal effect of income on expected (average) household food expenditure, holding all else constant, as shown in equation (2.3). If $E(e_i|x_i) \neq 0$ then x_i can be used to predict a nonzero value for e_i, which in turn will affect the value of y_i. In this case, β_2 will not capture all the effects of an income change, and the model cannot be interpreted as causal.

Another important consequence of the assumption of strict exogeneity is that it allows us to think of the econometric model as decomposing the dependent variable into two components: one that varies systematically as the values of the independent variable change and another that is random "noise." That is, the econometric model $y_i = \beta_1 + \beta_2 x_i + e_i$ can be broken into two parts: $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ and the random error, e_i. Thus

$$y_i = \beta_1 + \beta_2 x_i + e_i = E(y_i|x_i) + e_i$$

The values of the dependent variable y_i vary systematically due to variation in the conditional mean $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, as the value of the explanatory variable changes, and the values of the dependent variable y_i vary randomly due to e_i. The conditional *pdf*s of e and y are identical except for their location, as shown in Figure 2.3. Two values of food expenditure y_1 and y_2 for households with $x = \$1000$ of weekly income are shown in Figure 2.4 relative to their conditional mean. There will be variation in household expenditures on food from one household to another because of variations in tastes and preferences, and everything else. Some will spend more than the average value for households with the same income, and some will spend less. If we knew β_1 and β_2, then we could compute the conditional mean expenditure $E(y_i|x = 1000) = \beta_1 + \beta_2(1000)$ and also the value of the random errors e_1 and e_2. We never know β_1 and β_2 so we can never compute e_1 and e_2. What we are assuming, however, is that at each level of income x the average value of all that is represented by the random error is zero.

2.2.4 Random Error Variation

We have made the assumption that the conditional expectation of the random error term is zero, $E(e_i|x_i) = 0$. For the random error term we are interested in both its conditional mean,

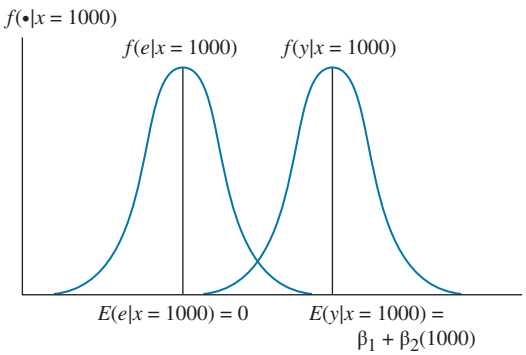


FIGURE 2.3 **Conditional probability densities for e and y.**

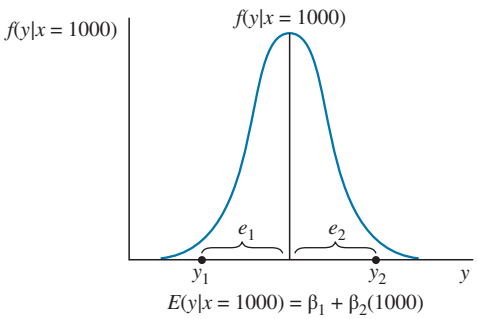


FIGURE 2.4 **The random error.**

or expected value, and its variance. Ideally, the **conditional variance** of the random error is constant.

$$\text{var}(e_i|x_i) = \sigma^2 \tag{2.4}$$

This is the **homoskedasticity** (also spelled homoscedasticity) assumption. At each x_i the variation of the random error component is the same. Assuming the population relationship $y_i = \beta_1 + \beta_2 x_i + e_i$ the conditional variance of the dependent variable is

$$\text{var}(y_i|x_i) = \text{var}(\beta_1 + \beta_2 x_i + e_i|x_i) = \text{var}(e_i|x_i) = \sigma^2$$

The simplification works because by conditioning on x_i we are treating it as if it is known and therefore not random. Given x_i the component $\beta_1 + \beta_2 x_i$ is not random, so the variance rule (P.14) applies.

This was an explicit assumption in Figure 2.1(b) where the *pdf*s $f(y|x = 1000)$ and $f(y|x = 2000)$ have the same variance, σ^2. If strict exogeneity holds, then the regression function is $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, as shown in Figure 2.2. The conditional distributions $f(y|x = 1000)$ and $f(y|x = 2000)$ are placed along the conditional mean function in Figure 2.5. In the household expenditure example, the idea is that for a particular level of household income x, the values of household food expenditure will vary randomly about the conditional mean due to the assumption that at each x the average value of the random error e is zero. Consequently, at each level of income, household food expenditures are centered on the regression function. The conditional homoskedasticity assumption implies that at each level of income the variation in food

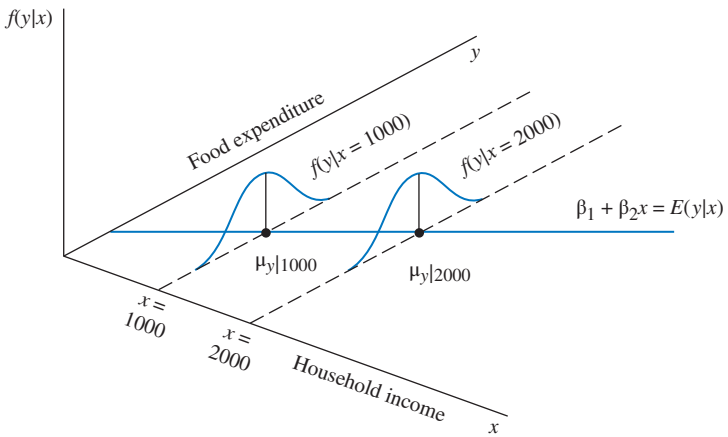


FIGURE 2.5 **The conditional probability density functions for y, food expenditure, at two levels of income.**

expenditure about its mean is the same. That means that at each and every level of income we are *equally* uncertain about how far food expenditures might fall from their mean value, $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Furthermore, this uncertainty does not depend on income or anything else. If this assumption is violated, and $\text{var}(e_i|x_i) \neq \sigma^2$, then the random errors are said to be **heteroskedastic**.

2.2.5 Variation in x

In a regression analysis, one of the objectives is to estimate $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i$. If we are to hope that a sample of data can be used to estimate the effects of changes in x, then we must observe some different values of the explanatory variable x in the sample. Intuitively, if we collect data **only** on households with income \$1000, we will not be able to measure the effect of changing income on the average value of food expenditure. Recall from elementary geometry that "it takes two points to determine a line." The minimum number of x-values in a sample of data that will allow us to proceed is two. You will find out in Section 2.4.4 that in fact the more different values of x, and the more variation they exhibit, the better our regression analysis will be.

2.2.6 Error Normality

In the discussion surrounding Figure 2.1, we explicitly made the assumption that food expenditures, given income, were normally distributed. In Figures 2.3–2.5, we implicitly made the assumption of conditionally normally distributed errors and dependent variable by drawing classically bell-shaped curves. It is not at all necessary for the random errors to be conditionally normal in order for regression analysis to "work." However, as you will discover in Chapter 3, when samples are small, it is advantageous for statistical inferences that the random errors, and dependent variable y, given each x-value, are normally distributed. The normal distribution has a long and interesting history,[3] as a little Internet searching will reveal. One argument for assuming regression errors are normally distributed is that they represent a collection of many different factors. The **Central Limit Theorem**, see Appendix C.3.4, says roughly that collections of many random factors tend toward having a normal distribution. In the context of the food expenditure model, if we consider that the random errors reflect tastes and preferences, it is entirely plausible that the random errors at each income level are normally distributed. When the assumption of conditionally normal errors is made, we write $e_i|x_i \sim N(0, \sigma^2)$ and also then $y_i|x_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$. It is a very strong assumption when it is made, and as mentioned it is not strictly speaking necessary, so we call it an *optional* assumption.

2.2.7 Generalizing the Exogeneity Assumption

So far we have assumed that the data pairs (y_i, x_i) have been drawn from a random sample and are *iid*. What happens if the sample values of the explanatory variable are correlated? And how might that happen?

A lack of independence occurs naturally when using financial or macroeconomic **time-series** data. Suppose we observe the monthly report on new housing starts, y_t, and the current 30-year fixed mortgage rate, x_t, and we postulate the model $y_t = \beta_1 + \beta_2 x_t + e_t$. The data (y_t, x_t) can be described as macroeconomic **time-series** data. In contrast to cross-section data where we have observations on a number of units (say households or firms or persons or countries) at a given point in time, with time-series data we have observations over time on a number of variables. It is customary to use a "t" subscript to denote time-series data and to use T to denote the sample size. In the data pairs (y_t, x_t), $t = 1, \ldots, T$, both y_t and x_t are random because we do not know the values

[3] For example, Stephen M. Stigler (1990) *The History of Statistics: The Measurement of Uncertainty, Reprint Edition*, Belknap Press, 73–76.

until they are observed. Furthermore, each of the data series is likely to be correlated across time. For example, the monthly fixed mortgage rate is likely to change slowly over time making the rate at time t correlated with the rate at time $t-1$. The assumption that the pairs (y_t, x_t) represent random *iid* draws from a probability distribution is not realistic. When considering the exogeneity assumption for this case, we need to be concerned not just with possible correlation between x_t and e_t, but also with possible correlation between e_t and every other value of the explanatory variable, namely, x_s, $s = 1, 2, \ldots, T$. If x_s is correlated with x_t, then it is possible that x_s (say, the mortgage rate in one month) may have an impact on y_t (say, housing starts in the next month). Since it is x_t, not x_s that appears in the equation $y_t = \beta_1 + \beta_2 x_t + e_t$, the effect of x_s will be included in e_t, implying $E(e_t|x_s) \neq 0$. We could use x_s to help predict the value of e_t. This possibility is ruled out when the pairs (y_t, x_t) are assumed to be independent. That is, independence of the pairs (y_t, x_t) **and** the assumption $E(e_t|x_t) = 0$ imply $E(e_t|x_s) = 0$ for all $s = 1, 2, \ldots, T$.

To extend the strict exogeneity assumption to models where the values of x are correlated, we need to assume $E(e_t|x_s) = 0$ for all $(t, s) = 1, 2, \ldots, T$. This means that we cannot predict the random error at time t, e_t, using any of the values of the explanatory variable. Or, in terms of our earlier notation, $E(e_i|x_j) = 0$ for all $(i, j) = 1, 2, \ldots, N$. To write this assumption in a more convenient form, we introduce the notation $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. That is, we are using $\mathbf{x}$ to denote all sample observations on the explanatory variable. Then, a more general way of writing the strict exogeneity assumption is $E(e_i|\mathbf{x}) = 0, i = 1, 2, \ldots, N$. From this assumption, we can also write $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ for $i = 1, 2, \ldots, N$. This assumption is discussed further in the context of alternative types of data in Section 2.10 and in Chapter 9. The assumption $E(e_i|\mathbf{x}) = 0$, $i = 1, 2, \ldots, N$, is a weaker assumption than assuming $E(e_i|x_i) = 0$ **and** that the pairs (y_i, x_i) are independent, and it enables us to derive a number of results for cases where different observations on x may be correlated as well as for the case where they are independent.

2.2.8 Error Correlation

In addition to possible correlations between a random error for one household (e_i) or one time period (e_t) being correlated with the value of an explanatory variable for another household (x_j) or time period (x_s), it is possible that there are correlations between the random error terms.

With cross-sectional data, data on households, individuals, or firms collected at one point in time, there may be a lack of statistical independence between random errors for individuals who are **spatially** connected. That is, suppose that we collect observations on two (or more) individuals who live in the same neighborhood. It is very plausible that there are similarities among people who live in a particular neighborhood. Neighbors can be expected to have similar incomes if the homes in a neighborhood are homogenous. Some suburban neighborhoods are popular because of green space and schools for young children, meaning households may have members similar in ages and interests. We might add a spatial component s to the error and say that the random errors $e_i(s)$ and $e_j(s)$ for the ith and jth households are possibly correlated because of their common location. Within a larger sample of data, there may be **clusters** of observations with correlated errors because of the spatial component.

In a time-series context, your author is writing these pages on the tenth anniversary of Hurricane Katrina, which devastated the U.S. Gulf Coast and the city of New Orleans, Louisiana, in particular. The impact of that shock did not just happen and then go away. The effect of that huge random event had an effect on housing and financial markets during August 2005, and also in September, October, and so on, to this day. Consequently, the random errors in the population relationship $y_t = \beta_1 + \beta_2 x_t + e_t$ are correlated over time, so that $\text{cov}(e_t, e_{t+1}) \neq 0$, $\text{cov}(e_t, e_{t+2}) \neq 0$, and so on. This is called **serial correlation**, or **autocorrelation**, in econometrics.

The starting point in regression analysis is to assume that there is no error correlation. In time-series models, we start by assuming $\text{cov}(e_t, e_s|\mathbf{x}) = 0$ for $t \neq s$, and for cross-sectional data we start by assuming $\text{cov}(e_i, e_j|\mathbf{x}) = 0$ for $i \neq j$. We will cope with failure of these assumptions in Chapter 9.

2.2.9 Summarizing the Assumptions

We summarize the starting assumptions of the simple regression model in a very general way. In our summary we use subscripts i and j but the assumptions are general, and apply equally to time-series data. If these assumptions hold, then regression analysis can successfully estimate the unknown population parameters β_1 and β_2 and we can claim that $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i = dE(y_i|x_i)/dx_i$ measures a causal effect. We begin our study of regression analysis and econometrics making these strong assumptions about the DGP. For future reference, the assumptions are named SR1–SR6, "SR" denoting "simple regression."

Econometrics is in large part devoted to handling data and models for which these assumptions **may not** hold, leading to modifications of usual methods for estimating β_1 and β_2, testing hypotheses, and predicting outcomes. In Chapters 2 and 3, we study the simple regression model under these, or similar, strong assumptions. In Chapter 4, we introduce modeling issues and diagnostic testing. In Chapter 5, we extend our model to **multiple regression analysis** with more than one explanatory variable. In Chapter 6, we treat modeling issues concerning the multiple regression model, and starting in Chapter 8 we address situations in which SR1–SR6 are violated in one way or another.

Assumptions of the Simple Linear Regression Model

SR1: Econometric Model All data pairs (y_i, x_i) collected from a population satisfy the relationship

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

SR2: Strict Exogeneity The conditional expected value of the random error e_i is zero. If $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, then

$$E(e_i|\mathbf{x}) = 0$$

If strict exogeneity holds, then the population regression function is

$$E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, N$$

and

$$y_i = E(y_i|\mathbf{x}) + e_i, \quad i = 1, \ldots, N$$

SR3: Conditional Homoskedasticity The conditional variance of the random error is constant.

$$\text{var}(e_i|\mathbf{x}) = \sigma^2$$

SR4: Conditionally Uncorrelated Errors The conditional covariance of random errors e_i and e_j is zero.

$$\text{cov}(e_i, e_j|\mathbf{x}) = 0 \quad \text{for} \quad i \neq j$$

SR5: Explanatory Variable Must Vary In a sample of data, x_i must take at least two different values.

SR6: Error Normality (*optional*) The conditional distribution of the random errors is normal.

$$e_i|\mathbf{x} \sim N(0, \sigma^2)$$

The random error e and the dependent variable y are both random variables, and as we have shown, the properties of one variable can be determined from the properties of the other. There is, however, one interesting difference between them: y is "observable" and e is "unobservable."

If the **regression parameters** β_1 and β_2 were *known*, then for any values y_i and x_i we could calculate $e_i = y_i - (\beta_1 + \beta_2 x_i)$. This is illustrated in Figure 2.4. Knowing the regression function $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ we could separate y_i into its systematic and random parts. However, β_1 and β_2 *are never known*, and it is impossible to calculate e_i.

What comprises the error term e? The random error e represents all factors affecting y other than x, or what we have called *everything else*. These factors cause individual observations y_i to differ from the conditional mean value $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$. In the food expenditure example, what factors can result in a difference between household expenditure per person y_i and its conditional mean $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$?

1. We have included income as the only explanatory variable in this model. Any *other* economic factors that affect expenditures on food are "collected" in the error term. Naturally, in any economic model, we want to include all the important and relevant explanatory variables in the model, so the error term e is a "storage bin" for unobservable and/or unimportant factors affecting household expenditures on food. As such, it adds noise that masks the relationship between x and y.
2. The error term e captures any approximation error that arises because the *linear* functional form we have assumed may be only an approximation to reality.
3. The error term captures any elements of random behavior that may be present in each individual. Knowing all the variables that influence a household's food expenditure might not be enough to perfectly predict expenditure. Unpredictable human behavior is also contained in e.

If we have omitted some important factor, or made any other serious **specification error**, then assumption SR2 $E(e_i|\mathbf{x}) = 0$ will be violated, which will have serious consequences.

2.3 Estimating the Regression Parameters

EXAMPLE 2.3 | Food Expenditure Model Data

The economic and econometric models we developed in the previous section are the basis for using a sample of data to *estimate* the intercept and slope parameters, β_1 and β_2. For illustration we examine typical data on household food expenditure and weekly income from a random sample of 40 households. Representative observations and summary statistics are given in Table 2.1. We control for household size by considering only three-person households. The values of y are weekly food expenditures for a three-person household, in dollars. Instead of measuring income in dollars, we measure it in units of $100, because a $1 increase in income has a numerically small effect on food expenditure. Consequently, for the first household, the reported income is $369 per week with weekly food expenditure of $115.22. For the 40th household, weekly income is $3340 and weekly food expenditure is $375.73. The complete data set of observations is in the data file *food*.

TABLE 2.1 Food Expenditure and Income Data

Observation (household)	Food Expenditure ($)	Weekly Income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary Statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. dev.	112.6752	6.8478

Remark

In this book, data files are referenced with a descriptive name in italics such as *food*. The actual files which are located at the book websites www.wiley.com/college/hill and www.principlesofeconometrics.com come in various formats and have an extension that denotes the format, for example, *food.dat*, *food.wf1, food.dta*, and so on. The corresponding data definition file is *food.def*.

We assume that the expenditure data in Table 2.1 satisfy the assumptions SR1–SR5. That is, we assume that the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ describes a population relationship and that the random error has conditional expected value zero. This implies that the conditional expected value of household food expenditure is a linear function of income. The conditional variance of y, which is the same as that of the random error e, is assumed constant, implying that we are equally uncertain about the relationship between y and x for all observations. Given **x** the values of y for different households are assumed uncorrelated with each other.

Given this theoretical model for explaining the sample observations on household food expenditure, the problem now is how to use the sample information in Table 2.1, specific values of y_i and x_i, to estimate the unknown regression parameters β_1 and β_2. These parameters represent the unknown intercept and slope coefficients for the food expenditure–income relationship. If we represent the 40 data points as (y_i, x_i), $i = 1, \ldots, N = 40$, and plot them, we obtain the **scatter diagram** in Figure 2.6.

Remark

It will be our notational convention to use i subscripts for cross-sectional data observations, with the number of sample observations being N. For time-series data observations, we use the subscript t and label the total number of observations T. In purely algebraic or generic situations, we may use one or the other.

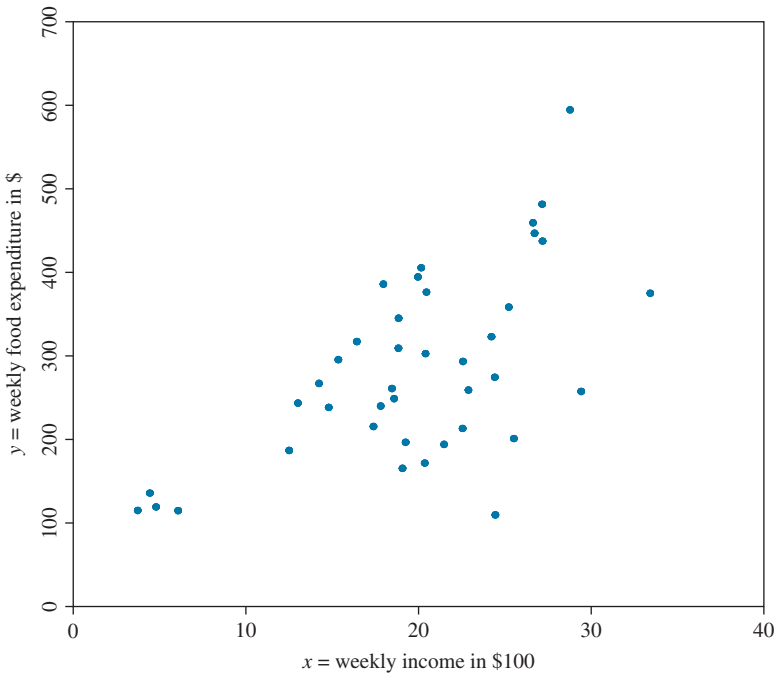


FIGURE 2.6 **Data for the food expenditure example.**

Our problem is to estimate the location of the mean expenditure line $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$. We would expect this line to be somewhere in the middle of all the data points since it represents population mean, or average, behavior. To estimate β_1 and β_2, we could simply draw a freehand line through the middle of the data and then measure the slope and intercept with a ruler. The problem with this method is that different people would draw different lines, and the lack of a formal criterion makes it difficult to assess the accuracy of the method. Another method is to draw a line from the expenditure at the smallest income level, observation $i = 1$, to the expenditure at the largest income level, $i = 40$. This approach does provide a formal rule. However, it may not be a very good rule because it ignores information on the exact position of the remaining 38 observations. It would be better if we could devise a rule that uses all the information from all the data points.

2.3.1 The Least Squares Principle

To estimate β_1 and β_2 we want a rule, or formula, that tells us how to make use of the sample observations. Many rules are possible, but the one that we will use is based on the **least squares principle**. This principle asserts that to fit a line to the data values we should make the sum of the squares of the vertical distances from each point to the line as small as possible. The distances are squared to prevent large positive distances from being canceled by large negative distances. This rule is arbitrary, but very effective, and is simply one way to describe a line that runs through the middle of the data. The intercept and slope of this line, the line that best fits the data using the least squares principle, are b_1 and b_2, the **least squares estimates** of β_1 and β_2. The fitted line itself is then

$$\hat{y}_i = b_1 + b_2 x_i \tag{2.5}$$

The vertical distances from each point to the fitted line are the **least squares residuals**. They are given by

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i \tag{2.6}$$

These residuals are depicted in Figure 2.7a.

Now suppose we fit another line, *any other line*, to the data. Denote the new line as

$$\hat{y}_i^* = b_1^* + b_2^* x_i$$

where b_1^* and b_2^* are any other intercept and slope values. The residuals for this line, $\hat{e}_i^* = y_i - \hat{y}_i^*$, are shown in Figure 2.7b. The least squares estimates b_1 and b_2 have the property that the sum of their squared residuals is *less than* the sum of squared residuals for *any* other line. That is, if

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2$$

is the sum of squared least squares residuals from (2.6) and

$$SSE^* = \sum_{i=1}^{N} \hat{e}_i^{*2} = \sum_{i=1}^{N} \left(y_i - \hat{y}_i^*\right)^2$$

is the sum of squared residuals based on any other estimates, then

$$SSE < SSE^*$$

no matter how the other line might be drawn through the data. The least squares principle says that the estimates b_1 and b_2 of β_1 and β_2 are the ones to use, since the line using them as intercept and slope fits the data best.

The problem is to find b_1 and b_2 in a convenient way. Given the sample observations on y and x, we want to find values for the unknown parameters β_1 and β_2 that minimize the "sum of squares" function

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} \left(y_i - \beta_1 - \beta_2 x_i\right)^2$$

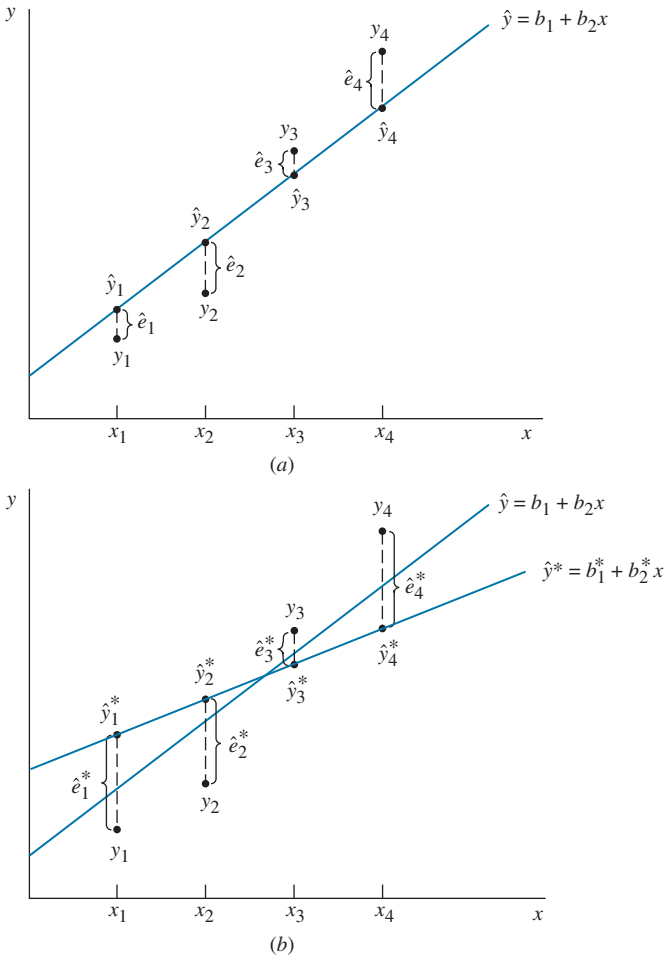


FIGURE 2.7 **(*a*) The relationship among y, $\hat{e}$, and the fitted regression line. (*b*) The residuals from another fitted line.**

This is a straightforward calculus problem, the details of which are given in Appendix 2A. The formulas for the least squares estimates of β_1 and β_2 that give the minimum of the sum of squared residuals are

The Ordinary Least Squares (OLS) Estimators

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.7}$$

$$b_1 = \bar{y} - b_2\bar{x} \tag{2.8}$$

where $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$ are the sample means of the observations on y and x.

We will call the estimators b_1 and b_2, given in equations (2.7) and (2.8), the **ordinary least squares estimators**. "Ordinary least squares" is abbreviated as **OLS**. These least squares estimators are called "ordinary," despite the fact that they are extraordinary, because these estimators are used day in and day out in many fields of research in a routine way, and to distinguish them

from other methods called **generalized least squares**, and **weighted least squares**, and **two-stage least squares**, all of which are introduced later in this book.

The formula for b_2 reveals why we had to assume [SR5] that in the sample x_i must take at least two different values. If $x_i = 5$, for example, for all observations, then b_2 in (2.7) is mathematically undefined and does not exist since its numerator and denominator are zero!

If we plug the sample values y_i and x_i into (2.7) and (2.8), then we obtain the least squares *estimates* of the intercept and slope parameters β_1 and β_2. It is interesting, however, and very important, that the formulas for b_1 and b_2 are perfectly general and can be used no matter what the sample values turn out to be. This should ring a bell. When the formulas for b_1 and b_2 are taken to be rules that are used whatever the sample data turn out to be, then b_1 and b_2 are random variables. When actual sample values are substituted into the formulas, we obtain numbers that are the observed values of random variables. To distinguish these two cases, we call the rules or general formulas for b_1 and b_2 the **least squares estimators**. We call the numbers obtained when the formulas are used with a particular sample **least squares estimates**.

- Least squares *estimators* are general formulas and are *random variables*.
- Least squares *estimates* are numbers that we obtain by applying the general formulas to the observed data.

The distinction between *estimators* and *estimates* is a fundamental concept that is essential to understand everything in the rest of this book.

EXAMPLE 2.4a | Estimates for the Food Expenditure Function

Using the least squares estimators (2.7) and (2.8), we can obtain the least squares estimates for the intercept and slope parameters β_1 and β_2 in the food expenditure example using the data in Table 2.1. From (2.7), we have

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

and from (2.8)

$$b_1 = \bar{y} - b_2\bar{x} = 283.5735 - (10.2096)(19.6048) = 83.4160$$

A convenient way to report the values for b_1 and b_2 is to write out the *estimated* or *fitted* regression line, with the estimates rounded appropriately:

$$\hat{y}_i = 83.42 + 10.21x_i$$

This line is graphed in Figure 2.8. The line's slope is 10.21, and its intercept, where it crosses the vertical axis, is 83.42. The least squares fitted line passes through the middle of the data in a very precise way, since one of the characteristics of the fitted line based on the least squares parameter estimates is that it passes through the point defined by the sample means, $(\bar{x}, \bar{y}) = (19.6048, 283.5735)$. This follows directly from rewriting (2.8) as $\bar{y} = b_1 + b_2\bar{x}$. Thus, the "point of the means" is a useful reference value in regression analysis.

Interpreting the Estimates

Once obtained, the least squares estimates are interpreted in the context of the economic model under consideration. The value $b_2 = 10.21$ is an estimate of β_2. Recall that x, weekly household income, is measured in \$100 units. The regression slope β_2 is the amount by which expected weekly expenditure on food per household increases when household weekly income increases by \$100. Thus, we estimate that if weekly household income goes up by \$100, expected weekly expenditure on food will increase by approximately \$10.21, holding all else constant. A supermarket executive with information on likely changes in the income and the number of households in an area could estimate that it will sell \$10.21 more per typical household per week for every \$100 increase in income. This is a very valuable piece of information for long-run planning.

Strictly speaking, the intercept estimate $b_1 = 83.42$ is an estimate of the expected weekly food expenditure for a household with zero income. In most economic models we must be very careful when interpreting the estimated intercept. The problem is that we often do not have any data points near $x = 0$, something that is true for the food expenditure data shown in Figure 2.8. If we have no observations in the region where income is zero, then our estimated relationship may not be a good approximation to reality in that region. So, although our estimated model suggests that a household with zero income is expected to spend \$83.42 per week on food, it might be risky to take this estimate literally. This is an issue that you should consider in each economic model that you estimate.

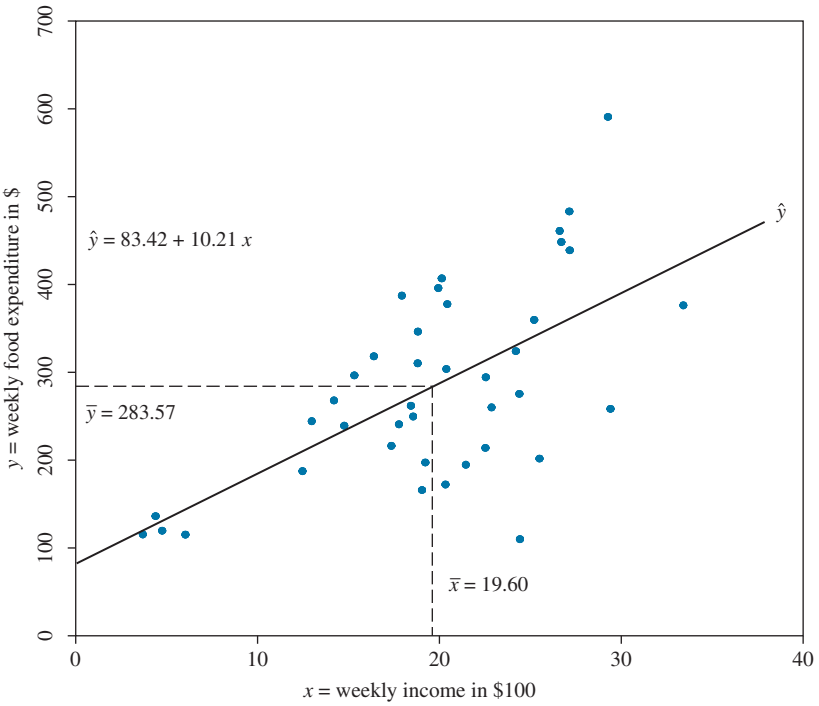


FIGURE 2.8 **The fitted regression.**

Elasticities Income elasticity is a useful way to characterize the responsiveness of consumer expenditure to changes in income. See Appendix A.2.2 for a discussion of elasticity calculations in a linear relationship. The **elasticity** of a variable y with respect to another variable x is

$$\varepsilon = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{100(\Delta y/y)}{100(\Delta x/x)} = \frac{\Delta y}{\Delta x}\cdot\frac{x}{y}$$

In the linear economic model given by (2.1), we have shown that

$$\beta_2 = \frac{\Delta E(y|\mathbf{x})}{\Delta x}$$

so the elasticity of mean expenditure with respect to income is

$$\varepsilon = \frac{\Delta E(y|\mathbf{x})}{\Delta x}\cdot\frac{x}{E(y|\mathbf{x})} = \beta_2\cdot\frac{x}{E(y|\mathbf{x})} \tag{2.9}$$

EXAMPLE 2.4b | Using the Estimates

To estimate this elasticity we replace β_2 by $b_2 = 10.21$. We must also replace "x" and "$E(y|\mathbf{x})$" by something, since in a linear model the elasticity is different on each point on the regression line. Most commonly, the elasticity is calculated at the "point of the means" $(\bar{x}, \bar{y}) = (19.60, 283.57)$ because it is a representative point on the regression line. If we calculate the income elasticity at the point of the means, we obtain

$$\hat{\varepsilon} = b_2\frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

This *estimated* income elasticity takes its usual interpretation. We estimate that a 1% increase in weekly household income will lead to a 0.71% increase in expected weekly household expenditure on food, when x and y take their sample mean values, $(\bar{x}, \bar{y}) = (19.60, 283.57)$. Since the estimated income elasticity is less than one, we would classify food as a "necessity" rather than a "luxury," which is consistent with what we would expect for an average household.

Prediction

The estimated equation can also be used for prediction or forecasting purposes. Suppose that we wanted to predict average weekly food expenditure for a household with a weekly income of $2000. This prediction is carried out by substituting $x = 20$ into our estimated equation to obtain

$$\hat{y}_i = 83.42 + 10.21x_i = 83.42 + 10.21(20) = 287.61$$

We *predict* that a household with a weekly income of $2000 will on average spend $287.61 per week on food.

Computer Output

Many different software packages can compute least squares estimates. Every software package's regression output looks different and uses different terminology to describe the output. Despite these differences, the various outputs provide the same basic information, which you should be able to locate and interpret. The matter is complicated somewhat by the fact that the packages also report various numbers whose meaning you may not know. For example, using the food expenditure data, the output from the software package EViews is shown in Figure 2.9.

In the EViews output, the parameter estimates are in the "Coefficient" column, with names "*C*," for constant term (the estimate b_1) and *INCOME* (the estimate b_2). Software programs typically name the estimates with the name of the variable as assigned in the computer program (we named our variable *INCOME*) and an abbreviation for "constant." The estimates that we report in the text are rounded to two significant digits. The other numbers that you can recognize at this time are $SSE = \sum \hat{e}_i^2 = 304505.2$, which is called "Sum squared resid," and the sample mean of y, $\bar{y} = \sum y_i/N = 283.5735$, which is called "Mean dependent var."

We leave discussion of the rest of the output until later.

Dependent Variable: *FOOD_EXP*
Method: Least Squares
Sample: 1 40
Included observations: 40

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000
R-squared	0.385002	Mean dependent var		283.5735
Adjusted R-squared	0.368818	S.D. dependent var		112.6752
S.E. of regression	89.51700	Akaike info criterion		11.87544
Sum squared resid	304505.2	Schwarz criterion		11.95988
Log likelihood	−235.5088	Hannan-Quinn criter		11.90597
F-statistic	23.78884	Durbin-Watson stat		1.893880
Prob(F-statistic)	0.000019			

FIGURE 2.9 **EViews regression output.**

2.3.2 Other Economic Models

We have used the household expenditure on food versus income relationship as an example to introduce the ideas of simple regression. The simple regression model can be applied to estimate

the parameters of many relationships in economics, business, and the social sciences. The applications of regression analysis are fascinating and useful. For example,

- If the hourly wage rate of electricians rises by 5%, how much will new house prices increase?
- If the cigarette tax increases by $1, how much additional revenue will be generated in the state of Louisiana?
- If the central banking authority raises interest rates by one-half a percentage point, how much will consumer borrowing fall within six months? How much will it fall within one year? What will happen to the unemployment rate in the months following the increase?
- If we increase funding on preschool education programs in 2018, what will be the effect on high school graduation rates in 2033? What will be the effect on the crime rate by juveniles in 2028 and subsequent years?

The range of applications spans economics and finance, as well as most disciplines in the social and physical sciences. Any time you ask **how much** a change in one variable will affect another variable, regression analysis is a potential tool.

Similarly, any time you wish to **predict** the value of one variable given the value of another then least squares regression is a tool to consider.

2.4 Assessing the Least Squares Estimators

Using the food expenditure data, we have estimated the parameters of the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the least squares formulas in (2.7) and (2.8). We obtained the least squares estimates $b_1 = 83.42$ and $b_2 = 10.21$. It is natural, but, as we shall argue, misguided, to ask the question "How good are these estimates?" This question is not answerable. We will never know the true values of the population parameters β_1 or β_2, so we cannot say how close $b_1 = 83.42$ and $b_2 = 10.21$ are to the true values. The least squares estimates are numbers that may or may not be close to the true parameter values, and we will never know.

Rather than asking about the quality of the estimates we will take a step back and examine the quality of the least squares estimation procedure. The motivation for this approach is this: if we were to collect another sample of data, by choosing another set of 40 households to survey, we would have obtained *different* estimates b_1 and b_2, even if we had carefully selected households with the same incomes as in the initial sample. This **sampling variation** is unavoidable. Different samples will yield different estimates because household food expenditures, y_i, $i = 1, \ldots, 40$, are random variables. Their values are not known until the sample is collected. Consequently, when viewed as an estimation procedure, b_1 and b_2 are also random variables, because their values depend on the random variable y. In this context, we call b_1 and b_2 the **least squares estimators**.

We can investigate the properties of the estimators b_1 and b_2, which are called their **sampling properties**, and deal with the following important questions:

1. If the least squares estimators b_1 and b_2 are random variables, then what are their expected values, variances, covariances, and probability distributions?
2. The least squares principle is only *one* way of using the data to obtain estimates of β_1 and β_2. How do the least squares estimators compare with other procedures that might be used, and how can we compare alternative estimators? For example, is there another estimator that has a higher probability of producing an estimate that is close to β_2?

We examine these questions in two steps to make things easier. In the first step, we investigate the properties of the least squares estimators conditional on the values of the explanatory variable in the sample. That is, conditional on **x**. Making the analysis conditional on **x** is equivalent to saying that, when we consider all possible samples, the household income values in the sample stay the

same from one sample to the next; only the random errors and food expenditure values change. This assumption is clearly not realistic but it simplifies the analysis. By conditioning on **x**, we are holding it constant, or fixed, meaning that we can treat the x-values as "not random."

In the second step, considered in Section 2.10, we return to the random sampling assumption and recognize that (y_i, x_i) data pairs are random, and randomly selecting households from a population leads to food expenditures and incomes that are random. However, even in this case and treating **x** as random, we will discover that most of our conclusions that treated **x** as nonrandom remain the same.

In either case, whether we make the analysis conditional on **x** or make the analysis general by treating **x** as random, the answers to the questions above depend critically on whether the assumptions SR1–SR5 are satisfied. In later chapters, we will discuss how to check whether the assumptions we make hold in a specific application, and what we might do if one or more assumptions are shown not to hold.

Remark

We will summarize the properties of the least squares estimators in the next several sections. "Proofs" of important results appear in the appendices to this chapter. In many ways, it is good to see these concepts in the context of a simpler problem before tackling them in the regression model. Appendix C covers the topics in this chapter, and the next, in the familiar and algebraically easier problem of estimating the mean of a population.

2.4.1 The Estimator b_2

Formulas (2.7) and (2.8) are used to compute the least squares estimates b_1 and b_2. However, they are not well suited for examining theoretical properties of the estimators. In this section, we rewrite the formula for b_2 to facilitate its analysis. In (2.7), b_2 is given by

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

This is called the **deviation from the mean form** of the estimator because the data have their sample means subtracted. Using assumption SR1 and a bit of algebra (Appendix 2C), we can write b_2 as a **linear estimator**,

$$b_2 = \sum_{i=1}^{N} w_i y_i \tag{2.10}$$

where

$$w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} \tag{2.11}$$

The term w_i depends only on **x**. Because we are conditioning our analysis on **x**, the term w_i is treated as if it is a constant. We remind you that conditioning on **x** is equivalent to treating **x** as given, as in a controlled, repeatable experiment.

Any estimator that is a weighted average of y_i's, as in (2.10), is called a **linear estimator**. This is an important classification that we will speak more of later. Then, with yet more algebra (Appendix 2D), we can express b_2 in a theoretically convenient way,

$$b_2 = \beta_2 + \sum w_i e_i \tag{2.12}$$

where e_i is the random error in the linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. This formula is not useful for computations, because it depends on β_2, which we do not know, and on the e_i's,

which are unobservable. However, for understanding the sampling properties of the least squares estimator, (2.12) is very useful.

2.4.2 The Expected Values of b_1 and b_2

The OLS estimator b_2 is a random variable since its value is unknown until a sample is collected. What we will show is that if our model assumptions hold, then $E(b_2|\mathbf{x}) = \beta_2$; that is, given $\mathbf{x}$ the expected value of b_2 is equal to the true parameter β_2. When the expected value of *any* estimator of a parameter equals the true parameter value, then that estimator is **unbiased**. Since $E(b_2|\mathbf{x}) = \beta_2$, the least squares estimator b_2 given $\mathbf{x}$ is an unbiased estimator of β_2. In Section 2.10, we will show that the least squares estimator b_2 is **unconditionally unbiased** also, $E(b_2) = \beta_2$. The intuitive meaning of unbiasedness comes from the sampling interpretation of mathematical expectation. Recognize that one sample of size N is just one of many samples that we could have been selected. If the formula for b_2 is used to estimate β_2 in each of those possible samples, then, if our assumptions are valid, the average value of the estimates b_2 obtained from all possible samples will be β_2.

We will show that this result is true so that we can illustrate the part played by the assumptions of the linear regression model. In (2.12), what parts are random? The parameter β_2 is not random. It is a population parameter we are trying to estimate. Conditional on $\mathbf{x}$ we can treat x_i as if it is not random. Then, conditional on $\mathbf{x}$, w_i is not random either, as it depends only on the values of x_i. The only random factors in (2.12) are the random error terms e_i. We can find the conditional expected value of b_2 using the fact that the expected value of a sum is the sum of the expected values:

$$\begin{aligned} E(b_2|\mathbf{x}) &= E(\beta_2 + \sum w_i e_i|\mathbf{x}) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N|\mathbf{x}) \\ &= E(\beta_2) + E(w_1 e_1|\mathbf{x}) + E(w_2 e_2|\mathbf{x}) + \cdots + E(w_N e_N|\mathbf{x}) \\ &= \beta_2 + \sum E(w_i e_i|\mathbf{x}) \\ &= \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 \end{aligned} \tag{2.13}$$

The rules of expected values are fully discussed in the Probability Primer, Section P.5, and Appendix B.1.1. In the last line of (2.13), we use two assumptions. First, $E(w_i e_i|\mathbf{x}) = w_i E(e_i|\mathbf{x})$ because conditional on $\mathbf{x}$ the terms w_i are not random, and constants can be factored out of expected values. Second, we have relied on the assumption that $E(e_i|\mathbf{x}) = 0$. Actually, if $E(e_i|\mathbf{x}) = c$, where c is any constant value, such as 3, then $E(b_2|\mathbf{x}) = \beta_2$. Given $\mathbf{x}$, the OLS estimator b_2 is an **unbiased estimator** of the regression parameter β_2. On the other hand, if $E(e_i|\mathbf{x}) \neq 0$ and it depends on $\mathbf{x}$ in some way, then b_2 is a **biased estimator** of β_2. One leading case in which the assumption $E(e_i|\mathbf{x}) = 0$ fails is due to **omitted variables**. Recall that e_i contains everything else affecting y_i other than x_i. If we have omitted anything that is important and that is correlated with $\mathbf{x}$ then we would expect that $E(e_i|\mathbf{x}) \neq 0$ and $E(b_2|\mathbf{x}) \neq \beta_2$. In Chapter 6 we discuss this **omitted variables bias**. Here we have shown that conditional on $\mathbf{x}$, and under SR1–SR5, the least squares estimator is linear and unbiased. In Section 2.10, we show that $E(b_2) = \beta_2$ without conditioning on $\mathbf{x}$.

The unbiasedness of the estimator b_2 is an important sampling property. On average, over all possible samples from the population, the least squares estimator is "correct," on average, and this is one desirable property of an estimator. This statistical property by itself does not mean that b_2 is a good estimator of β_2, but it is part of the story. The unbiasedness property is related to what happens in all possible samples of data from the same population. The fact that b_2 is unbiased does not imply *anything* about what might happen *in just one sample*. An individual estimate (a number) b_2 may be near to, or far from, β_2. Since β_2 is *never* known we will never know, given

one sample, whether our estimate is "close" to β_2 or not. Thus, the estimate $b_2 = 10.21$ may be close to β_2 or not.

The least squares estimator b_1 of β_1 is also an unbiased estimator, and $E(b_1|\mathbf{x}) = \beta_1$ if the model assumptions hold.

2.4.3 Sampling Variation

To illustrate how the concept of unbiased estimation relates to sampling variation, we present in Table 2.2 least squares estimates of the food expenditure model from 10 hypothetical random samples (data file *table2_2*) of size $N = 40$ from the same population with the same incomes as the households given in Table 2.1. Note the variability of the least squares parameter estimates from sample to sample. This **sampling variation** is due to the fact that we obtain 40 *different* households in each sample, and their weekly food expenditure varies randomly.

The property of unbiasedness is about the *average* values of b_1 and b_2 if used in all possible samples of the same size drawn from the same population. The average value of b_1 in these 10 samples is $\bar{b}_1 = 96.11$. The average value of b_2 is $\bar{b}_2 = 8.70$. If we took the averages of estimates from more samples, these averages would approach the true parameter values β_1 and β_2. Unbiasedness does not say that an estimate from any one sample is close to the true parameter value, and thus we cannot say that an *estimate* is unbiased. We can say that the least squares estimation procedure (or the least squares estimator) is unbiased.

2.4.4 The Variances and Covariance of b_1 and b_2

Table 2.2 shows that the least squares estimates of β_1 and β_2 vary from sample to sample. Understanding this variability is a key to assessing the reliability and sampling precision of an estimator. We now obtain the variances and covariance of the estimators b_1 and b_2. Before presenting the expressions for the variances and covariance, let us consider why they are important to know. The variance of the random variable b_2 is the average of the squared distances between the possible values of the random variable and its mean, which we now know is $E(b_2|\mathbf{x}) = \beta_2$. The conditional variance of b_2 is defined as

$$\text{var}(b_2|\mathbf{x}) = E\left\{\left[b_2 - E(b_2|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\}$$

TABLE 2.2 Estimates from 10 Hypothetical Samples

Sample	b_1	b_2
1	93.64	8.24
2	91.62	8.90
3	126.76	6.59
4	55.98	11.23
5	87.26	9.14
6	122.55	6.80
7	91.95	9.84
8	72.48	10.50
9	90.34	8.75
10	128.55	6.99

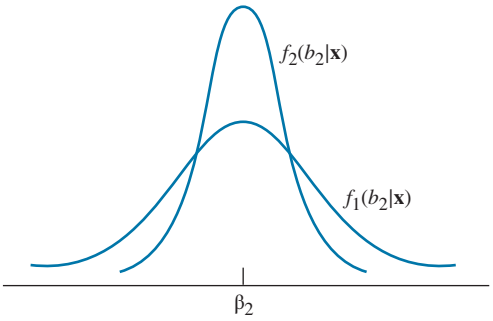


FIGURE 2.10 **Two possible probability density functions for b_2.**

It measures the spread of the probability distribution of b_2. In Figure 2.10 are graphs of two possible probability distributions of b_2, $f_1(b_2|\mathbf{x})$ and $f_2(b_2|\mathbf{x})$, that have the same mean value but different variances.

The *pdf* $f_2(b_2|\mathbf{x})$ has a smaller variance than $f_1(b_2|\mathbf{x})$. Given a choice, we are interested in estimator precision and would prefer that b_2 have the *pdf* $f_2(b_2|\mathbf{x})$, rather than $f_1(b_2|\mathbf{x})$. With the distribution $f_2(b_2|\mathbf{x})$, the probability is more concentrated around the true parameter value β_2 giving, relative to $f_1(b_2|\mathbf{x})$, a higher probability of getting an estimate that is close to β_2. Remember, getting an estimate close to β_2 is a primary objective of regression analysis.

The variance of an estimator measures the *precision* of the estimator in the sense that it tells us how much the estimates can vary from sample to sample. Consequently, we often refer to the **sampling variance** or **sampling precision** of an estimator. The smaller the variance of an estimator is, the greater the sampling precision of that estimator. One estimator is more precise than another estimator if its sampling variance is less than that of the other estimator.

We will now present and discuss the conditional variances and covariance of b_1 and b_2. Appendix 2E contains the derivation of the variance of the least squares estimator b_2. If the regression model assumptions SR1–SR5 are correct (assumption SR6 is not required), then the variances and covariance of b_1 and b_2 are

$$\text{var}(b_1|\mathbf{x}) = \sigma^2 \left[\frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2} \right] \tag{2.14}$$

$$\text{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2.15}$$

$$\text{cov}(b_1, b_2|\mathbf{x}) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \tag{2.16}$$

At the beginning of this section we said that for unbiased estimators, smaller variances are better than larger variances. Let us consider the factors that affect the variances and covariance in (2.14)–(2.16).

1. The variance of the random error term, σ^2, appears in each of the expressions. It reflects the dispersion of the values y about their expected value $E(y|\mathbf{x})$. The greater the variance σ^2, the greater is that dispersion, and the greater is the uncertainty about where the values of y fall relative to their conditional mean $E(y|\mathbf{x})$. When σ^2 is larger, the information we have about β_1 and β_2 is less precise. In Figure 2.5, the variance is reflected in the spread of the probability distributions $f(y|x)$. The *larger* the variance term σ^2, the *greater is* the uncertainty in the statistical model, and the *larger* the variances and covariance of the least squares estimators.

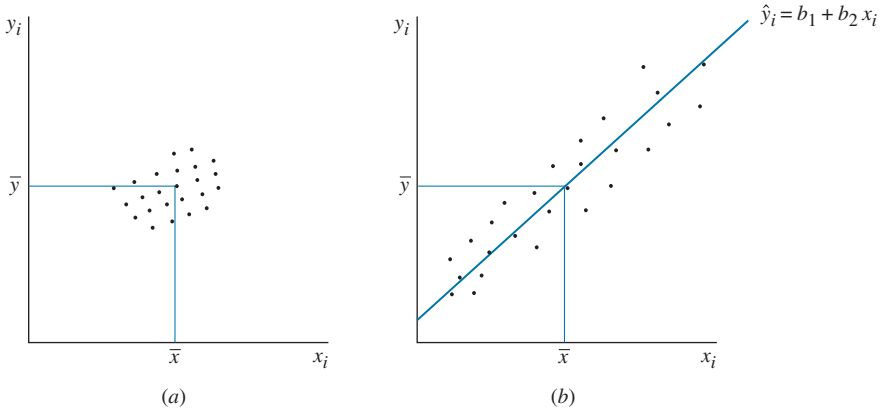


FIGURE 2.11 **The influence of variation in the explanatory variable x on precision of estimation: (*a*) low x variation, low precision: (*b*) high x variation, high precision.**

2. The sum of squares of the values of x about their sample mean, $\sum(x_i - \bar{x})^2$, appears in each of the variances and in the covariance. This expression measures how *spread out* about their mean are the sample values of the independent or explanatory variable x. The more they are spread out, the larger the sum of squares. The less they are spread out, the smaller the sum of squares. You may recognize this sum of squares as the numerator of the sample variance of the x-values. See Appendix C.4. The *larger* the sum of squares, $\sum(x_i - \bar{x})^2$, the *smaller* the conditional variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters. The intuition behind this is demonstrated in Figure 2.11. Panel (b) is a data scatter in which the values of x are widely spread out along the x-axis. In panel (a), the data are "bunched." Which data scatter would you prefer given the task of fitting a line by hand? Pretty clearly, the data in panel (b) do a better job of determining where the least squares line must fall, because they are more spread out along the x-axis.
3. The larger the sample size N, the *smaller* the variances and covariance of the least squares estimators; it is better to have *more* sample data than *less*. The sample size N appears in each of the variances and covariance because each of the sums consists of N terms. Also, N appears explicitly in $\text{var}(b_1|\mathbf{x})$. The sum of squares term $\sum(x_i - \bar{x})^2$ gets larger as N increases because each of the terms in the sum is positive or zero (being zero if x happens to equal its sample mean value for an observation). Consequently, as N gets larger, both $\text{var}(b_2|\mathbf{x})$ and $\text{cov}(b_1, b_2|\mathbf{x})$ get smaller, since the sum of squares appears in their denominator. The sums in the numerator and denominator of $\text{var}(b_1|\mathbf{x})$ both get larger as N gets larger and offset one another, leaving the N in the denominator as the dominant term, ensuring that $\text{var}(b_1|\mathbf{x})$ also gets smaller as N gets larger.
4. The term $\sum x_i^2$ appears in $\text{var}(b_1|\mathbf{x})$. The larger this term is, the larger the variance of the least squares estimator b_1. Why is this so? Recall that the intercept parameter β_1 is the expected value of y given that $x = 0$. The farther our data are from $x = 0$, the more difficult it is to interpret β_1, as in the food expenditure example, and the more difficult it is to accurately estimate β_1. The term $\sum x_i^2$ measures the squared distance of the data from the origin, $x = 0$. If the values of x are near zero, then $\sum x_i^2$ will be small, and this will reduce $\text{var}(b_1|\mathbf{x})$. But if the values of x are large in magnitude, either positive or negative, the term $\sum x_i^2$ will be large and $\text{var}(b_1)$ will be larger, other things being equal.
5. The sample mean of the x-values appears in $\text{cov}(b_1, b_2|\mathbf{x})$. The absolute magnitude of the covariance *increases* with an increase in magnitude of the sample mean $\bar{x}$, and the covariance has a *sign* opposite to that of $\bar{x}$. The reasoning here can be seen from Figure 2.11. In panel (b)

the least squares fitted line must pass through the point of the means. Given a fitted line through the data, imagine the effect of increasing the estimated slope b_2. Since the line must pass through the point of the means, the effect must be to lower the point where the line hits the vertical axis, implying a reduced intercept estimate b_1. Thus, when the sample mean is positive, as shown in Figure 2.11, there is a negative covariance between the least squares estimators of the slope and intercept.

2.5 The Gauss–Markov Theorem

What can we say about the least squares estimators b_1 and b_2 so far?

- The estimators are perfectly general. Formulas (2.7) and (2.8) can be used to estimate the unknown parameters β_1 and β_2 in the simple linear regression model, no matter what the data turn out to be. Consequently, viewed in this way, the least squares estimators b_1 and b_2 are random variables.
- The least squares estimators are *linear* estimators, as defined in (2.10). Both b_1 and b_2 can be written as weighted averages of the y_i values.
- If assumptions SR1–SR5 hold, then the least squares estimators are conditionally *unbiased*. This means that $E(b_1|\mathbf{x}) = \beta_1$ and $E(b_2|\mathbf{x}) = \beta_2$.
- Given $\mathbf{x}$ we have expressions for the variances of b_1 and b_2 and their covariance. Furthermore, we have argued that for any unbiased estimator, having a smaller variance is better, as this implies we have a higher chance of obtaining an estimate close to the true parameter value.

Now we will state and discuss the famous **Gauss–Markov theorem**, which is proven in Appendix 2F.

Gauss–Markov Theorem:

Given $\mathbf{x}$ and under the assumptions SR1–SR5 of the linear regression model, the estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are the **best linear unbiased estimators (BLUE)** of β_1 and β_2.

Let us clarify what the Gauss–Markov theorem does, and does not, say.

1. The estimators b_1 and b_2 are "best" when compared to similar estimators, those that are linear and unbiased. The theorem does *not* say that b_1 and b_2 are the best of all *possible* estimators.
2. The estimators b_1 and b_2 are best within their class because they have the minimum variance. When comparing two linear and unbiased estimators, we *always* want to use the one with the smaller variance, since that estimation rule gives us the higher probability of obtaining an estimate that is close to the true parameter value.
3. In order for the Gauss–Markov theorem to hold, assumptions SR1–SR5 must be true. If any of these assumptions are *not* true, then b_1 and b_2 are *not* the best linear unbiased estimators of β_1 and β_2.
4. The Gauss–Markov theorem does *not* depend on the assumption of normality (assumption SR6).
5. In the simple linear regression model, if we want to use a linear and unbiased estimator, then we have to do no more searching. The estimators b_1 and b_2 are the ones to use. This explains

why we are studying these estimators (we would not have you study *bad* estimation rules, would we?) and why they are so widely used in research, not only in economics but in all social and physical sciences as well.

6. The Gauss–Markov theorem applies to the least squares estimators. It *does not* apply to the least squares *estimates* from a single sample.

The results we have presented so far treat **x** as given. In Section 2.10 we show that the Gauss–Markov theorem also holds in general, and it does not depend on a specific **x**.

2.6 The Probability Distributions of the Least Squares Estimators

The properties of the least squares estimators that we have developed so far do not depend in any way on the normality assumption SR6. If we also make this assumption, that the random errors e_i are normally distributed, with mean zero and variance σ^2, then the conditional probability distributions of the least squares estimators are also normal. This conclusion is obtained in two steps. First, given **x** and based on assumption SR1, if e_i is normal then so is y_i. Second, the least squares estimators are linear estimators of the form $b_2 = \sum w_i y_i$. Given **x** this weighted sum of normal random variables is also normally distributed. Consequently, *if* we make the normality assumption (assumption SR6 about the error term), and treat **x** as given, then the least squares estimators are normally distributed:

$$b_1|\mathbf{x} \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum(x_i - \bar{x})^2}\right) \tag{2.17}$$

$$b_2|\mathbf{x} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) \tag{2.18}$$

As you will see in Chapter 3, the normality of the least squares estimators is of great importance in many aspects of statistical inference.

What if the errors are not normally distributed? Can we say anything about the probability distribution of the least squares estimators? The answer is, sometimes, yes.

A Central Limit Theorem:

If assumptions SR1–SR5 hold, and if the sample size N is **sufficiently large**, then the least squares estimators have a distribution that approximates the normal distributions shown in (2.17) and (2.18).

The million-dollar question is "How large is sufficiently large?" The answer is that there is no specific number. The reason for this vague and unsatisfying answer is that "how large" depends on many factors, such as what the distributions of the random errors look like (are they smooth? symmetric? skewed?) and what the x_i values are like. In the simple regression model, some would say that $N = 30$ is sufficiently large. Others would say that $N = 50$ would be a more reasonable number. The bottom line is, however, that these are rules of thumb and that the meaning of "sufficiently large" will change from problem to problem. Nevertheless, for better or worse, this *large sample*, or **asymptotic**, result is frequently invoked in regression analysis. This important result is an application of a central limit theorem, like the one discussed in Appendix C.3.4. If you are not familiar with this important theorem, you may want to review it now.

2.7 Estimating the Variance of the Error Term

The variance of the random error term, σ^2, is the one unknown parameter of the simple linear regression model that remains to be estimated.

The conditional variance of the random error e_i is

$$\text{var}(e_i|\mathbf{x}) = \sigma^2 = E\left\{\left[e_i - E(e_i|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\} = E(e_i^2|\mathbf{x})$$

if the assumption $E(e_i|\mathbf{x}) = 0$ is correct. Since the "expectation" is an average value, we might consider estimating σ^2 as the average of the squared errors

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N}$$

This formula is unfortunately of no use since the random errors e_i are *unobservable*! However, although the random errors themselves are unknown, we do have an analog to them—namely, the least squares residuals. Recall that the random errors are

$$e_i = y_i - \beta_1 - \beta_2 x_i$$

From (2.6) the least squares residuals are obtained by replacing the unknown parameters by their least squares estimates:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

It seems reasonable to replace the random errors e_i by their analogs, the least squares residuals, so that

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N}$$

This estimator, though quite satisfactory in large samples, is a *biased* estimator of σ^2. But there is a simple modification that produces an unbiased estimator:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} \tag{2.19}$$

The 2 that is subtracted in the denominator is the number of *regression parameters* (β_1, β_2) in the model, and this subtraction makes the estimator $\hat{\sigma}^2$ unbiased, so that $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$.

2.7.1 Estimating the Variances and Covariance of the Least Squares Estimators

Having an unbiased estimator of the error variance means we can *estimate* the conditional variances of the least squares estimators b_1 and b_2 and the covariance between them. Replace the unknown error variance σ^2 in (2.14)–(2.16) by $\hat{\sigma}^2$ to obtain

$$\widehat{\text{var}}(b_1|\mathbf{x}) = \hat{\sigma}^2 \left[\frac{\sum x_i^2}{N\sum(x_i - \bar{x})^2}\right] \tag{2.20}$$

$$\widehat{\text{var}}(b_2|\mathbf{x}) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \tag{2.21}$$

$$\widehat{\text{cov}}(b_1, b_2|\mathbf{x}) = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum(x_i - \bar{x})^2}\right] \tag{2.22}$$

The square roots of the estimated variances are the "standard errors" of b_1 and b_2. These quantities are used in hypothesis testing and confidence intervals. They are denoted as $\text{se}(b_1)$ and $\text{se}(b_2)$

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1|\mathbf{x})} \tag{2.23}$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} \tag{2.24}$$

EXAMPLE 2.5 | Calculations for the Food Expenditure Data

Let us make some calculations using the food expenditure data. The least squares estimates of the parameters in the food expenditure model are shown in Figure 2.9. First, we will compute the least squares residuals from (2.6) and use them to calculate the estimate of the error variance in (2.19). In Table 2.3 are the least squares residuals for the first five households in Table 2.1.

TABLE 2.3 Least Squares Residuals

x	y	$\hat{y}$	$\hat{e} = y - \hat{y}$
3.69	115.22	121.09	−5.87
4.39	135.98	128.24	7.74
4.75	119.34	131.91	−12.57
6.03	114.96	144.98	−30.02
12.47	187.05	210.73	−23.68

Recall that we have estimated that for the food expenditure data the fitted least squares regression line is $\hat{y} = 83.42 + 10.21x$. For each observation, we compute the least squares residual $\hat{e}_i = y_i - \hat{y}_i$. Using the residuals for all $N = 40$ observations, we estimate the error variance to be

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{304505.2}{38} = 8013.29$$

The numerator, 304505.2, is the sum of squared least squares residuals, reported as "Sum squared resid" in Figure 2.9. The denominator is the number of sample observations, $N = 40$, minus the number of estimated regression parameters, 2; the quantity $N - 2 = 38$ is often called the **degrees of freedom** for reasons that will be explained in Chapter 3. In Figure 2.9, the value $\hat{\sigma}^2$ is not reported. Instead, EViews software reports $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{8013.29} = 89.517$, labeled "S.E. of regression," which stands for "standard error of the regression."

It is typical for software not to report the estimated variances and covariance unless requested. However, all software packages automatically report the standard errors. For example, in the EViews output shown in Figure 2.9 the column labeled "Std. Error" contains $\text{se}(b_1) = 43.410$ and $\text{se}(b_2) = 2.093$. The entry called "S.D. dependent var" is the sample standard deviation of y, that is, $\left[\sum(y_i - \bar{y})^2/(N-1)\right]^{1/2} = 112.6752$.

The full set of estimated variances and covariances for a regression is usually obtained by a simple computer command, or option, depending on the software being used. They are arrayed in a rectangular array, or matrix, with variances on the diagonal and covariances in the "off-diagonal" positions.

$$\begin{bmatrix} \widehat{\text{var}}(b_1|\mathbf{x}) & \widehat{\text{cov}}(b_1, b_2|\mathbf{x}) \\ \widehat{\text{cov}}(b_1, b_2|\mathbf{x}) & \widehat{\text{var}}(b_2|\mathbf{x}) \end{bmatrix}$$

For the food expenditure data, the estimated covariance matrix of the least squares estimators is

	C	*INCOME*
C	1884.442	−85.90316
INCOME	−85.90316	4.381752

where C stands for the "constant term," which is the estimated intercept parameter in the regression, or b_1; similarly, the software reports the variable name *INCOME* for the column relating to the estimated slope b_2. Thus

$$\widehat{\text{var}}(b_1|\mathbf{x}) = 1884.442, \quad \widehat{\text{var}}(b_2|\mathbf{x}) = 4.381752,$$
$$\widehat{\text{cov}}(b_1, b_2|\mathbf{x}) = -85.90316$$

The standard errors are

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1|\mathbf{x})} = \sqrt{1884.442} = 43.410$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} = \sqrt{4.381752} = 2.093$$

These values will be used extensively in Chapter 3.

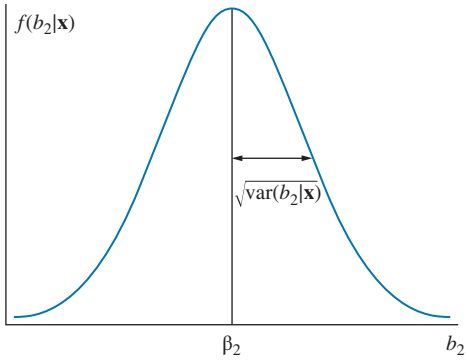


FIGURE 2.12 **The conditional probability density function of the least squares estimator b_2.**

2.7.2 Interpreting the Standard Errors

The standard errors of b_1 and b_2 are measures of the **sampling variability** of the least squares estimates b_1 and b_2 in **repeated samples**. As illustrated in Table 2.2, when we collect different samples of data, the parameter estimates change from sample to sample. The estimators b_1 and b_2 are general formulas that are used whatever the sample data turn out to be. That is, the estimators are random variables. As such, they have probability distributions, means, and variances. In particular, if assumption SR6 holds, and the random error terms e_i are normally distributed, then $b_2|\mathbf{x} \sim N\left(\beta_2, \text{var}(b_2|\mathbf{x}) = \sigma^2/\sum(x_i - \bar{x})^2\right)$. This *pdf* $f(b_2|\mathbf{x})$ is shown in Figure 2.12.

The estimator variance, $\text{var}(b_2|\mathbf{x})$, or, its square root $\sigma_{b_2} = \sqrt{\text{var}(b_2|\mathbf{x})}$, which we might call the true standard deviation of b_2, measures the sampling variation of the estimates b_2 and determines the width of the *pdf* in Figure 2.12. The bigger σ_{b_2} is the more variation in the least squares estimates b_2 we see from sample to sample. If σ_{b_2} is large, then the estimates might change a great deal from sample to sample. The parameter σ_{b_2} would be a valuable number to know, because if it were large relative to the parameter β_2 we would know that the least squares estimator is not precise, and the estimate that we obtain may be far from the true value β_2 that we are trying to estimate. On the other hand, if σ_{b_2} is small relative to the parameter β_2, we know that the least squares estimate will fall near β_2 with high probability. Recall that for the normal distribution, 99.9% of values fall within the range of three standard deviations from the mean, so that 99.9% of the least squares estimates will fall in the range $\beta_2 - 3\sigma_{b_2}$ to $\beta_2 + 3\sigma_{b_2}$.

To put this in another context, in Table 2.2 we report estimates from 10 samples of data. We noted in Section 2.4.3 that the average values of those estimates are $\bar{b}_1 = 96.11$ and $\bar{b}_2 = 8.70$. The question we address with the standard error is "How much variation about their means do the estimates exhibit from sample to sample?" For those 10 samples, the sample standard deviations are std. dev.$(b_1) = 23.61$ and std. dev.$(b_2) = 1.58$. What we would **really like** is the values of the standard deviations for a **very large** number of samples. Then we would know how much variation the least squares estimates exhibit from sample to sample. Unfortunately, we do not have a large number of samples, and because we do not know the true value of the variance of the error term σ^2 we cannot know the true value of σ_{b_2}.

Then what do we do? We estimate σ^2, and then estimate σ_{b_2} using

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} = \sqrt{\frac{\hat{\sigma}^2}{\Sigma(x_i - \bar{x})^2}}$$

The standard error of b_2 is thus an estimate of what the standard deviation of many estimates b_2 would be in a very large number of samples and is an indicator of the width of the *pdf* of b_2 shown in Figure 2.12. Using our one sample of data, *food*, the standard error of b_2 is 2.093, as shown in

the computer output in Figure 2.9. This value is reasonably close to std. dev. $(b_2) = 1.58$ from the 10 samples in Table 2.2. To put this to a further test, in Appendix 2H, we perform a simulation experiment, called a **Monte Carlo experiment**, in which we create many artificial samples to demonstrate the properties of the least squares estimator and how well $\text{se}(b_2)$ reflects the true sampling variation in the estimates.

2.8 Estimating Nonlinear Relationships

The world is not linear. Economic variables are not always related by straight-line relationships; in fact, many economic relationships are represented by curved lines and are said to display **curvilinear** forms. Fortunately, the simple linear regression model $y = \beta_1 + \beta_2 x + e$ is much more flexible than it looks at first glance, because the variables y and x can be transformations, involving logarithms, squares, cubes, or reciprocals, of the basic economic variables, or they can be **indicator variables** that take only the values zero and one. Including these possibilities means the simple linear regression model can be used to account for **nonlinear relationships** between variables.[4]

Nonlinear relationships can sometimes be anticipated. Consider a model from real estate economics in which the price (*PRICE*) of a house is related to the house size measured in square feet (*SQFT*). As a starting point, we might consider the linear relationship

$$PRICE = \beta_1 + \beta_2 SQFT + e \tag{2.25}$$

In this model, β_2 measures the increase in expected price given an additional square foot of living area. In the linear specification, the expected price per additional square foot is constant. However, it may be reasonable to assume that larger and more expensive homes have a higher value for an additional square foot of living area than smaller, less expensive homes. How can we build this idea into our model? We will illustrate the use of two approaches: first, a **quadratic** equation in which the explanatory variable is $SQFT^2$; and second, a **log-linear** equation in which the dependent variable is ln(*PRICE*). In each case, we will find that the slope of the relationship between *PRICE* and *SQFT* is not constant, but changes from point to point.

2.8.1 Quadratic Functions

The quadratic function $y = a + bx^2$ is a parabola.[5] The y-intercept is a. The shape of the curve is determined by b; if $b > 0$, then the curve is U-shaped; and if $b < 0$, then the curve has an inverted-U shape. The slope of the function is given by the derivative[6] $dy/dx = 2bx$, which changes as x changes. The elasticity or the percentage change in y given a 1% change in x is $\varepsilon = slope \times x/y = 2bx^2/y$. If a and b are greater than zero, the curve resembles Figure 2.13.

2.8.2 Using a Quadratic Model

A **quadratic model** for house prices includes the **squared** value of *SQFT*, giving

$$PRICE = \alpha_1 + \alpha_2 SQFT^2 + e \tag{2.26}$$

This is a simple regression model, $y = \alpha_1 + \alpha_2 x + e$, with $y = PRICE$ and $x = SQFT^2$. Here, we switch from using β to denote the parameters to using α, because the parameters of (2.26) are not comparable to the parameters of (2.25). In (2.25) β_2 is a slope, but α_2 is not a slope. Because

[4]The term linear in "linear regression" means that the parameters are not transformed in any way. In a linear regression model, the parameters must not be raised to powers or transformed, so expressions like $\beta_1\beta_2$ or $\beta_2^{\beta_1}$ are not permitted.

[5]This is a special case of the more general quadratic function $y = a + bx + cx^2$.

[6]See Appendix A.3.1, Derivative Rules 1–5.

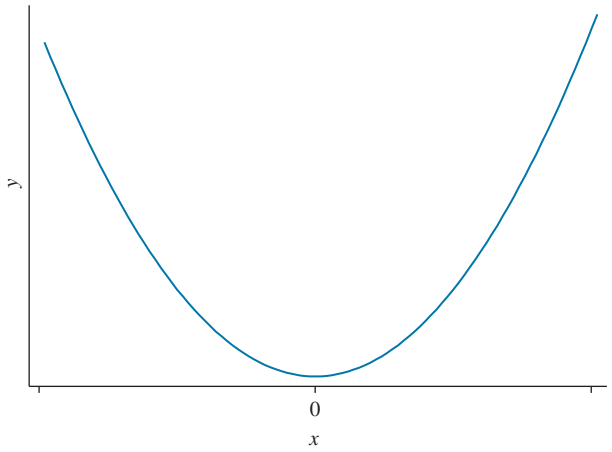


FIGURE 2.13 **A quadratic function, $y = a + bx^2$.**

$SQFT > 0$, the house price model will resemble the right side of the curve in Figure 2.13. Using ^ to denote estimated values, the least squares estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$, of α_1 and α_2, are calculated using the estimators in (2.7) and (2.8), just as earlier. The fitted equation is $\widehat{PRICE} = \hat{\alpha}_1 + \hat{\alpha}_2 SQFT^2$. It has slope

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2\hat{\alpha}_2 SQFT \tag{2.27}$$

If $\hat{\alpha}_2 > 0$, then larger houses will have larger slope, and a larger estimated price per additional square foot.

EXAMPLE 2.6 | Baton Rouge House Data

The data file *br* contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. Using these data, the estimated quadratic equation is $\widehat{PRICE} = 55776.56 + 0.0154SQFT^2$. The data scatter and fitted quadratic relationship are shown in Figure 2.14.

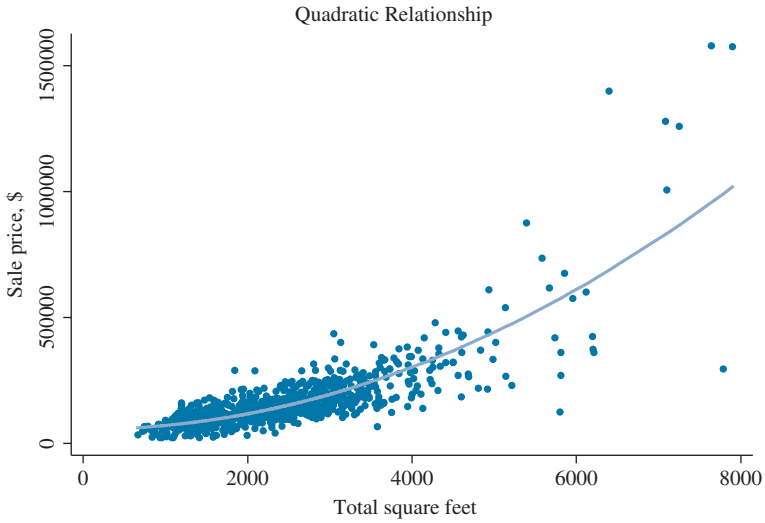


FIGURE 2.14 **A fitted quadratic relationship.**

The estimated slope is $\widehat{slope} = 2(0.0154)\,SQFT$ (estimated price per additional square foot), which for a 2000-square-foot house is \$61.69, for a 4000-square-foot house is \$123.37, and for a 6000-square-foot house is \$185.05. The elasticity of house price with respect to house size is the percentage increase in estimated price given a 1% increase in house size. Like the slope, the elasticity changes at each point. In our example

$$\hat{\varepsilon} = \widehat{slope} \times \frac{SQFT}{PRICE} = \left(2\hat{\alpha}_2 SQFT\right) \times \frac{SQFT}{PRICE}$$

To compute an estimate, we must select values for *SQFT* and *PRICE* on the fitted relationship. That is, we choose a value for *SQFT* and choose for price the corresponding fitted value $\widehat{PRICE}$. For houses of 2000, 4000, and 6000 square feet, the estimated elasticities are 1.05 [using $\widehat{PRICE}$ = \$117,461.77], 1.63 [using $\widehat{PRICE}$ = \$302,517.39], and 1.82 [using $\widehat{PRICE}$ = \$610,943.42], respectively. For a 2000-square-foot house, we estimate that a 1% increase in house size will increase price by 1.05%.

2.8.3 A Log-Linear Function

The log-linear equation $\ln(y) = a + bx$ has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side. Both its slope and elasticity change at each point and are the same sign as b. Using the antilogarithm, we see that $\exp[\ln(y)] = y = \exp(a + bx)$, so that the log-linear function is an exponential function. The function requires $y > 0$. The slope[7] at any point is $dy/dx = \exp(a + bx) \times b = by$, which for $b > 0$ means that the marginal effect increases for larger values of y. An economist might say that this function is increasing at an increasing rate, as shown in Figure 2.15.

The elasticity, the percentage change in y given a 1% increase in x, at a point on this curve is $\varepsilon = slope \times x/y = bx$.

Using the slope expression, we can solve for a **semi-elasticity**, which tells us the percentage change in y given a one-unit increase in x. Divide both sides of the slope dy/dx by y, then multiply by 100 to obtain

$$\eta = \frac{100(dy/y)}{dx} = 100b \tag{2.28}$$

In this expression, the numerator $100(dy/y)$ is the percentage change in y; dx represents the change in x. If $dx = 1$, then a one-unit change in x leads to a $100b$ percentage change in y. This interpretation can sometimes be quite handy.

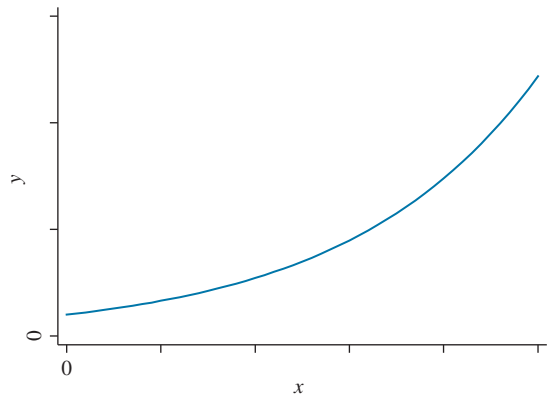


FIGURE 2.15 **A log-linear function.**

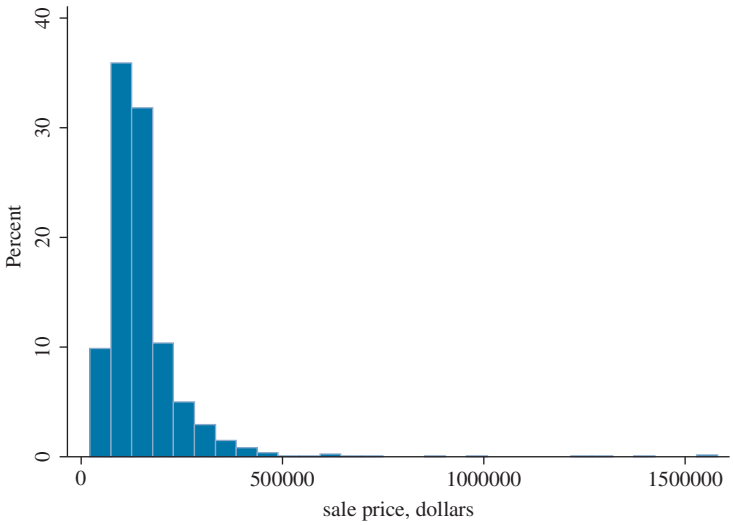
[7] See Appendix A.3.1, Derivative Rule 7.

2.8.4 Using a Log-Linear Model

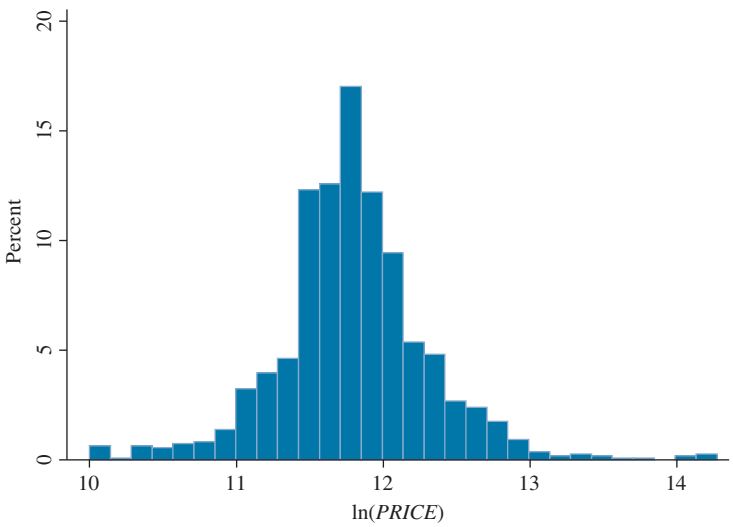
The use of logarithms is very common in economic modeling. The **log-linear model** uses the logarithm of a variable as the dependent variable, and an independent, explanatory variable, that is not transformed, such as[8]

$$\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e \tag{2.29}$$

What effects does this have? First, the logarithmic transformation can regularize data that is skewed with a long tail to the right. In Figure 2.16(a), we show the histogram of *PRICE* and in Figure 2.16(b) the histogram of ln(*PRICE*). The median house price in this sample is $130,000,



(*a*)



(*b*)

FIGURE 2.16 **(*a*) Histogram of *PRICE*. (*b*) Histogram of ln(*PRICE*).**

[8]Once again we use different symbols for the parameters of this model, γ_1 and γ_2, as a reminder that these parameters are not directly comparable to β's in (2.25) or α's in (2.26).

and 95% of house prices are below $315,000, but there are 24 houses out of the 1080 with prices above $500,000, and an extreme value of $1,580,000. The extremely skewed distribution of *PRICE* becomes more symmetric, if not bell-shaped, after taking the logarithm. Many economic variables, including prices, incomes, and wages, have skewed distributions, and the use of logarithms in models for such variables is common.

Second, using a log-linear model allows us to fit regression curves like that shown in Figure 2.15.

EXAMPLE 2.7 | Baton Rouge House Data, Log-Linear Model

Using the Baton Rouge data, the fitted log-linear model is

$$\widehat{\ln(PRICE)} = 10.8386 + 0.0004113\ SQFT$$

To obtain predicted price, take the antilogarithm,[9] which is the exponential function

$$\widehat{PRICE} = \exp\left[\widehat{\ln(PRICE)}\right] = \exp(10.8386 + 0.0004113\ SQFT)$$

The fitted value of *PRICE* is shown in Figure 2.17.

The slope of the log-linear model is

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = \hat{\gamma}_2 \widehat{PRICE} = 0.0004113\ \widehat{PRICE}$$

For a house with a predicted *PRICE* of $100,000, the estimated increase in *PRICE* for an additional square foot of house area is $41.13, and for a house with a predicted *PRICE* of $500,000, the estimated increase in *PRICE* for an additional square foot of house area is $205.63. The estimated elasticity is $\hat{\varepsilon} = \hat{\gamma}_2\ SQFT = 0.0004113\ SQFT$. For a house with 2000 square feet, the estimated elasticity is 0.823: a 1% increase in house size is estimated to increase selling price by 0.823%. For a house with 4000 square feet, the estimated elasticity is 1.645: a 1% increase in house size is estimated to increase selling price by 1.645%. Using the "semi-elasticity" defined in equation (2.28), we can say that, for a one-square-foot increase in size, we estimate a price increase of 0.04%. Or, perhaps more usefully, we estimate that a 100-square-foot increase will increase price by approximately 4%.

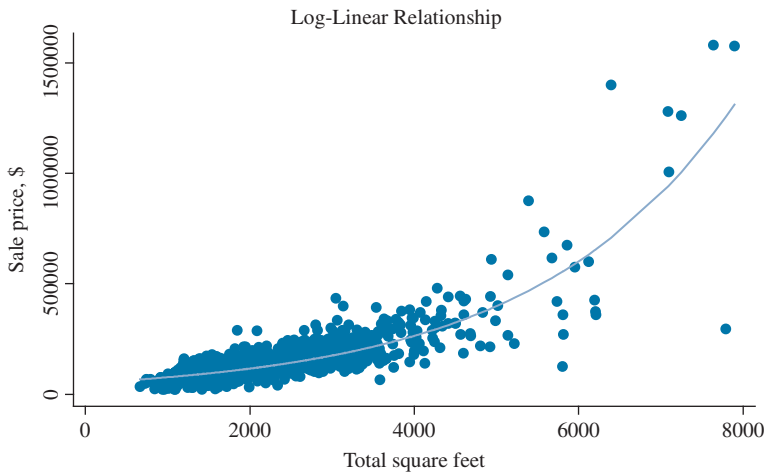


FIGURE 2.17 **The fitted log-linear model.**

[9] In Chapter 4 we present an improved predictor for this model.

2.8.5 Choosing a Functional Form

For the Baton Rouge house price data, should we use the quadratic functional form or the log-linear functional form? This is not an easy question. Economic theory tells us that house price should be related to the size of the house, and perhaps that larger, more expensive homes have a higher price per additional square foot of living area. But economic theory does not tell us what the exact algebraic form of the relationship should be. We should do our best to choose a functional form that is consistent with economic theory, that fits the data well, and that is such that the assumptions of the regression model are satisfied. In real-world problems, it is sometimes difficult to achieve all these goals. Furthermore, we will never truly know the correct functional relationship, no matter how many years we study econometrics. The truth is out there, but we will never know it. In applications of econometrics, we must simply do the best we can to choose a satisfactory functional form. At this point, we mention one dimension of the problem used for evaluating models with the same dependent variable. By comparing the sum of squared residuals (*SSE*) of alternative models, or, equivalently, $\hat{\sigma}^2$ or $\hat{\sigma}$, we can choose the model that is a better fit to the data. Smaller values of these quantities mean a smaller sum of squared residuals and a better model fit. This comparison is **not** valid for comparing models with dependent variables y and $\ln(y)$, or when other aspects of the models are different. We study the choice among functions like these further in Chapter 4.

2.9 Regression with Indicator Variables

An indicator variable is a binary variable that takes the values zero or one; it is used to represent a nonquantitative characteristic, such as gender, race, or location. For example, in the data file *utown.dot* we have a sample of 1,000 observations on house prices (*PRICE*, in thousands of dollars) in two neighborhoods. One neighborhood is near a major university and called University Town. Another similar neighborhood, called Golden Oaks, is a few miles away from the university. The indicator variable of interest is

$$UTOWN = \begin{cases} 1 & \text{house is in University Town} \\ 0 & \text{house is in Golden Oaks} \end{cases}$$

The histograms of the prices in these two neighborhoods, shown in Figure 2.18, are revealing. The mean of the distribution of house prices in University Town appears to be larger than the mean of the distribution of house prices from Golden Oaks. The sample mean of the 519 house prices in University Town is 277.2416 thousand dollars, whereas the sample mean of the 481 Golden Oaks houses is 215.7325 thousand dollars.

If we include *UTOWN* in a regression model as an explanatory variable, what do we have? The simple regression model is

$$PRICE = \beta_1 + \beta_2 UTOWN + e$$

If the regression assumptions SR1–SR5 hold, then the least squares estimators in (2.7) and (2.8) can be used to estimate the unknown parameters β_1 and β_2.

When an indicator variable is used in a regression, it is important to write out the regression function for the different values of the indicator variable.

$$E(PRICE|UTOWN) = \beta_1 + \beta_2 UTOWN = \begin{cases} \beta_1 + \beta_2 & \text{if } UTOWN = 1 \\ \beta_1 & \text{if } UTOWN = 0 \end{cases}$$

In this case, we find that the "regression function" reduces to a model that implies that the population mean house prices in the two subdivisions are different. The parameter β_2 is not a slope

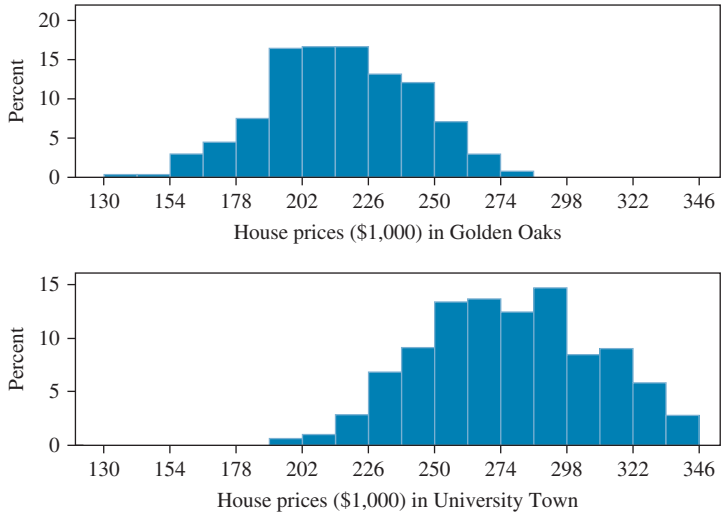


FIGURE 2.18 **Distributions of house prices.**

in this model. Here β_2 is the difference between the population means for house prices in the two neighborhoods. The expected price in University Town is $\beta_1 + \beta_2$, and the expected price in Golden Oaks is β_1. In our model, there are no factors other than location affecting price, and the indicator variable splits the observations into two populations.

The estimated regression is

$$\widehat{PRICE} = b_1 + b_2 UTOWN = 215.7325 + 61.5091\, UTOWN$$

$$= \begin{cases} 277.2416 & \text{if } UTOWN = 1 \\ 215.7325 & \text{if } UTOWN = 0 \end{cases}$$

We see that the estimated price for the houses in University Town is \$277,241.60, which is also the sample mean of the house prices in University Town. The estimated price for houses outside University Town is \$215,732.50, which is the sample mean of house prices in Golden Oaks.

In the regression model approach, we estimate the regression intercept β_1, which is the expected price for houses in Golden Oaks, where $UTOWN = 0$, and the parameter β_2, which is the difference between the population means for house prices in the two neighborhoods. The least squares estimators b_1 and b_2 in this indicator variable regression can be shown to be

$$b_1 = \overline{PRICE}_{\text{Golden Oaks}}$$

$$b_2 = \overline{PRICE}_{\text{University Town}} - \overline{PRICE}_{\text{Golden Oaks}}$$

where $\overline{PRICE}_{\text{Golden Oaks}}$ is the sample mean (average) price of houses in Golden Oaks and $\overline{PRICE}_{\text{University Town}}$ is the sample mean price of houses from University Town.

In the simple regression model, an indicator variable on the right-hand side gives us a way to estimate the differences between population means. This is a common problem in statistics, and the direct approach using samples means is discussed in Appendix C.7.2. Indicator variables are used in regression analysis very frequently in many creative ways. See Chapter 7 for a full discussion.

2.10 The Independent Variable[10]

Earlier in this chapter we specified a number of assumptions for the simple regression model and then used these assumptions to derive some properties of the least squares estimators of the coefficients in the model. In the household food expenditure example, we assumed a DGP where pairs (y_i, x_i) are randomly drawn from some population. We then went on to make a strict exogeneity assumption $E(e_i|\mathbf{x}) = 0$ to accommodate other types of DGPs. Using this and other assumptions, we derived properties of the least squares estimator conditional on the sample values $\mathbf{x}$. In this section, we say more about different possible DGPs, explore their implications for the assumptions of the simple regression model, and investigate how the properties of the least squares estimator change, if at all, when we no longer condition on $\mathbf{x}$.

Our regression model $y = \beta_1 + \beta_2 x + e$ has five components, three of which are unobservable: β_1, β_2, and e. The two observable components are y the random outcome, or dependent variable, and x the explanatory, independent variable. Is this explanatory variable random or not and why does it matter? We address these questions in this section.

How do we obtain values for the observable pair of variables (y, x)? In an experimental DGP, a scientist under carefully controlled conditions specifies the values of x, performs an experiment, and observes the outcomes y. For example, an agronomist might vary the number of pounds of pesticide spread per acre of cropland and observe the resulting yield. In this case, the independent variable, pounds of pesticide, is in fact an *independent* factor and not random. It is fixed. It is not affected by random influences and the treatment can be replicated time and time again. Laboratory and other controlled experiments can claim that the values of the independent variable are fixed. In the world of economics and business, there are few examples of laboratory and controlled experiments.[11] One exception is retail sales. Merchants display the prices of goods and services and observe consumer purchases. The merchant controls the prices, store displays, advertising and the shopping environment. In this case, we can argue that x, the price of a product in a retail store, is fixed and not random; it is *given*. When x is fixed and not random, the idea of repeated experimental trials makes intuitive sense. The sampling properties of the least squares estimators are a summary of how the estimators perform under a series of controlled experiments with fixed values for the independent variables. We have shown that the least squares estimator is the best linear unbiased estimator, given $\mathbf{x}$, and we have variance equations (2.14) and (2.15) that describe how much variation the estimates exhibit from sample to sample.

In the next three sections, we treat cases in which x-values are random. Each of these cases represents a different type of DGP. We start with the strongest assumption about random-x and then look at weaker cases.

2.10.1 Random and Independent x

Suppose our agronomist takes another strategy, using a random number between 0 and 100 to determine the amount of pesticide applied to a given acre of land. In this case, x is random, as its value is unknown until it is randomly selected. Why might a scientist use this approach? Well, no one could imply that such an experiment was rigged to produce a particular outcome. It is a "fair" experiment because the scientist keeps "hands off" the controls. What are the sampling properties of the least squares estimator in this setting? Is the least squares estimator the best, linear unbiased estimator in this case?

[10] This section contains a more advanced discussion of the assumptions of the simple regression model.

[11] Economists understand the benefits of controlled experiments. The field of experimental economics has grown tremendously in the past 20 years. Also, there have been some social experiments. One example is Tennessee's Project STAR that examined the consequences on school children of having small classes rather than larger ones. This example is explored further in Chapter 7.5.

In order to answer these questions, we make explicit that x is *statistically independent* of the error term e. The assumptions for the **independent random-x model** (IRX) are as follows:

Assumptions of the Independent Random-x Linear Regression Model

IRX1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

IRX2: The random error has mean zero, $E(e_i) = 0$.

IRX3: The random error has constant variance, $\text{var}(e_i) = \sigma^2$.

IRX4: The random errors e_i and e_j for any two observations are uncorrelated, $\text{cov}(e_i, e_j) = 0$.

IRX5: The random errors $e_1, e_2, \ldots, e_N$ are statistically independent of $x_1, \ldots, x_N$, and x_i takes at least two different values.

IRX6: $e_i \sim N(0, \sigma^2)$.

Compare the assumptions IRX2, IRX3, and IRX4 with the initial assumptions about the simple regression model, SR2, SR3, and SR4. You will note that conditioning on $\mathbf{x}$ has disappeared. The reason is because when x-values and random errors e are statistically independent $E(e_i|x_j) = E(e_i) = 0$, $\text{var}(e_i|x_j) = \text{var}(e_i) = \sigma^2$ and $\text{cov}(e_i, e_j|\mathbf{x}) = \text{cov}(e_i, e_j) = 0$. Refer back to the Probability Primer Sections P.6.1 and P.6.2 for a discussion of why conditioning has no effect on the expected value and variance of statistically independent random variables. Also, it is extremely important to recognize that "i" and "j" simply represent different data observations that may be cross-sectional data or time-series data. What we say applies to both types of data.

The least squares estimators b_1 and b_2 are the best linear unbiased estimators of β_1 and β_2 if assumptions IRX1–IRX5 hold. These results are derived in Appendix 2G.2. The one apparent change is that an "expected value" appears in the formulas for the estimator variances. For example,

$$\text{var}(b_2) = \sigma^2 E\left[\frac{1}{\sum(x_i - \bar{x})^2}\right]$$

We must take the expected value of the term involving x. In practice, this actually changes nothing, because we estimate the variance in the usual way.

$$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}$$

The estimator of the error variance remains $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-2)$ and all the usual interpretations remain the same. Thus, the computational aspects of least squares regression do not change. What has changed is our understanding of the DGP. Furthermore, if IRX6 holds then, conditional on $\mathbf{x}$, the least squares estimators have normal distributions.[12]

As we will see in Chapter 3, procedures for inference, namely interval estimators and hypothesis tests, will work in this independent random-x model the same way as in a fixed-x model. And, thanks to the central limit theorem, cited in Section 2.6, it will still be true that in *large samples* the least squares estimator has an approximate normal distribution whether x is fixed or random. This will be explored further in Chapter 5.

[12] If we do not condition on $\mathbf{x}$, no longer treating it as fixed and given, the exact distribution of the least squares estimator is not normal and is in fact unknown. Equation (2.12) shows that b_2 is a complicated combination of x's and random errors, e. Even if we know the distributions of x and e the product of random variables w_i and e_i has an unknown distribution.

2.10.2 Random and Strictly Exogenous x

Statistical independence between x_i and e_j, for all values of i and j (which may denote time-series or cross-sectional observations) is a very strong assumption and most likely only suitable in experimental situations. A weaker assumption is that the explanatory variable x is **strictly exogenous**. The phrases "strictly exogenous" and "strict exogeneity" refer to a particular technical, statistical assumption. You have no doubt heard the term **exogenous** before in your principles of economics classes. For example, in a supply and demand model, we know that the equilibrium price and quantity in a competitive market are jointly determined by the forces of supply and demand. Price and quantity are **endogenous** variables that are determined within the equilibrium system. However, we know that consumer income affects the demand equation. If income increases, the demand for a normal good increases. Income is not determined within the equilibrium system that determines equilibrium price and quantity; it is determined outside this market and is said to be **exogenous**. The exogenous variable income affects market demand, but market demand does not affect consumer income. In regression analysis models, the independent, explanatory variable x is also termed an **exogenous variable** because its variation affects the outcome variable y, but there is no reverse causality; changes in y have no effect on x.

Because interrelationships among economic variables and forces can be complex, we wish to be very precise about exogenous explanatory variables. The independent variable x is **strictly exogenous** if $E(e_i|x_j) = 0$ for all values of i and j, or equivalently, $E(e_i|x_1, x_2, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$. This is exactly assumption SR2. If $i = 3$, for example, then $E(e_3|x_1) = 0$, and $E(e_3|x_3) = 0$, and $E(e_3|x_7) = 0$. The conditional expectation of the ith error term e_i is zero given *any and all* x_j. If it will help you remember them, relabel SR1–SR6 as SEX1–SEX6, where SEX stands for "strictly exogenous-x." Let the phrase "simple regression is sexy" remind you that ***S***trictly ***E***xogenous-***X*** is the baseline regression assumption.

What are the properties of the least squares estimator under the assumption of strict exogeneity? They are the same as in the case of statistical independence between all x_j and e_i. The least squares estimators are the best linear unbiased estimators of the regression parameters. These results are proved in Appendix 2G.3. This is a nice finding because while still strong, strict exogeneity is less strong than assuming x and e are statistically independent. Furthermore, if the errors are normally distributed, then the least squares estimator $b_2|\mathbf{x}$ has a normal distribution.

The Implications of Strict Exogeneity

Strict exogeneity implies quite a bit. If x is strictly exogenous, then the least squares estimator works the way we want it to and no fancier or more difficult estimators are required. Life is simple. If, on the other hand, strict exogeneity does not hold, then econometric analysis becomes more complicated, which, unfortunately, is often the case. How can we tell if the technical, statistical assumption called "strict exogeneity" holds? The only sure way is to perform a controlled experiment in which x is fixed in repeated samples or chosen randomly as described in Section 2.10.1. For most economic analyses, such experiments are impossible or too expensive.

Are there perhaps some statistical tests that can be used to check for strict exogeneity? The answer is yes, but using statistics it is much easier to determine if something is probably false rather than to argue that it is true. The common practice is to check that the implications of strict exogeneity are true. If these implications don't seem to be true, either based on economic logic or statistical tests, then we will conclude that strict exogeneity does not hold and deal with the consequences, making life more difficult. The two direct implications of strict exogeneity, $E(e_i|x_1, x_2, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$, derived in Appendix 2G.1, are as follows:

Implication 1: $E(e_i) = 0$. The "average" of all factors omitted from the regression model is zero.

Implication 2: $\text{cov}(x_i, e_j) = 0$. There is no correlation between the omitted factors associated with observation j and the value of the explanatory variable for observation i.

If x satisfies the strict exogeneity condition, then $E(e_i) = 0$ and $\text{cov}(x_i, e_j) = 0$. If either of these implications is *not true*, then x is *not strictly exogenous*.

Can we check Implication 1: $E(e_i) = 0$? Is the average of all omitted factors equal to zero? In practice, this usually reduces to the question "Have I omitted anything important from the model?" If you have it is likely to be because you didn't know it was important (weak economic theory) or because, while you know it is an important factor (such as an individual's average lifetime income or an individual's perseverance in the face of adversity), it cannot be easily or well measured. In any event, omitted variables damage the least squares estimator only when Implication 2 is violated. Consequently, Implication 2 draws the most attention.

Can we check Implication 2: $\text{cov}(x_i, e_j) = 0$? Yes, we can, and we show some statistical tests in Chapter 10. However, logical arguments, and thought experiments, should always come before any statistical tests. In some cases, we can anticipate the failure of strict exogeneity, as the following examples in models using *time-series* data illustrate. In these cases, we usually index the observations using the subscript t, so that x_t is the value of the explanatory variable at time t and e_s is the value of the random error in time period s. In this context, strict exogeneity would be expressed as $E(e_s|x_t) = 0$ for all s and t. The zero covariance implication of strict exogeneity is $\text{cov}(x_t, e_s) = 0$.

Example 1. Suppose that x_t represents a policy variable, perhaps public spending on roads and bridges in month or quarter t. If the area is "shocked" by a hurricane, tornado, or other natural disaster at time s, then some time later ($t > s$) we may very well expect public spending on roads and bridges to increase, not only for one time period but perhaps for many. Then, $\text{cov}(x_{t=s+1}, e_s) \neq 0$, $\text{cov}(x_{t=s+2}, e_s) \neq 0$, and so on. Strict exogeneity fails in this case because the shock to the error term, the natural disaster, is correlated with a subsequent change in the explanatory variable, public spending, implying $E(e_s|x_t) \neq 0$.

Example 2. Suppose the quarterly sales by a firm are related to its advertising expenditures. We might write $SALES_t = \beta_1 + \beta_2 ADVERT_t + e_t$. However, advertising expenditures at time t may depend on sales revenues in the same quarter during the previous year, at time $t - 4$. That is, $ADVERT_t = f(SALES_{t-4})$. Because $SALES_{t-4} = \beta_1 + \beta_2 ADVERT_{t-4} + e_{t-4}$, it follows that there will be a correlation, and covariance, between $ADVERT_t$ and e_{t-4}. Therefore, the strict exogeneity condition fails, and $E(e_{t-4}|ADVERT_t) \neq 0$. Note the similarities between this example and the first. The effect of a past error e_s is carried forward to affect a future value of the explanatory variable, x_t, $t > s$.

Example 3. Let U_t represent the unemployment rate in quarter t, and we suppose that it is affected by the governmental expenditures, G_t. The regression might be specified as $U_t = \beta_1 + \beta_2 G_t + e_t$. However, we might imagine that the unemployment rate in this quarter is affected by government spending in previous quarters, such as G_{t-1}. Because G_{t-1} is not included in the model specification, it makes up a portion of the error term, $e_t = f(G_{t-1})$. Furthermore, we expect that there is a strong positive correlation and covariance between government spending this quarter and in previous quarters, so that $\text{cov}(G_t, G_{t-1}) > 0$. This means that we can anticipate a correlation between the error term in time t and previous levels of government spending, so that $\text{cov}(e_t, G_{t-1}) \neq 0$. Therefore, $\text{cov}(e_t|G_t) \neq 0$ and the strict exogeneity assumption fails.

The implications of a failure of the strict exogeniety assumption for least squares estimation, and the introduction of weaker assumptions to accommodate situations like those in Examples 1–3, are considered in Chapters 5, 9, and 10.

2.10.3 Random Sampling

The food expenditure example we have carried through this chapter is another case in which the DGP leads to an x that is random. We randomly sampled a population and selected 40 households. These are cross-sectional data observations. For each household, we recorded their food expenditure (y_i) and income (x_i). Because both of these variables' values are unknown to us until they are observed, both the outcome variable y and the explanatory variable x are random.

The same questions are relevant. What are the sampling properties of the least squares estimator in this case? Is the least squares estimator the best, linear unbiased estimator?

Such survey data is collected by **random sampling** from a population. Survey methodology is an important area of statistics. Public opinion surveys, market research surveys, government surveys, and censuses are all examples of collecting survey data. Several important ones are carried out by the U.S. Bureau of Labor Statistics (BLS).[13] The idea is to collect data pairs (y_i, x_i) in such a way that the ith pair [the "Smith" household] is statistically independent of the jth pair [the "Jones" household]. This ensures that x_j is statistically independent of e_i if $i \neq j$. Then, the strict exogeneity assumption reduces to concern about a possible relationship between x_i and e_i. If the conditional expectation $E(e_i|x_i) = 0$, then x is strictly exogenous, and the implications are $E(e_i) = 0$ and $\text{cov}(x_i, e_i) = 0$. Note also that if we assume that the data pairs are independent, then we no longer need make the separate assumption that the errors are uncorrelated.

What are the properties of the least squares estimator under these assumptions? They are the same as in the cases of statistical independence between all x_j and e_i (Section 2.10.1) and strict exogeneity in the general sense (Section 2.10.2). The least squares estimators are the best linear unbiased estimators of the regression parameters and conditional on **x** they have a normal distribution if SR6 (or IRX6) holds.

One final idea associated with random sampling is that the data pairs, (y_i, x_i), $i = 1, \ldots, N$, have the same joint *pdf*, $f(y, x)$. In this case, the data pairs are independent and identically distributed, *iid*. In statistics, the phrase **random sample** implies that the data are ***iid***. This is a reasonable assumption if all the data pairs are collected from the same population.

When discussing examples of the implications of strict exogeneity, we showed how the strict exogeneity assumption can be violated when using time-series data if there is correlation between e_s and a future or past value x_t ($t \neq s$). For an example of how strict exogeneity fails with random sampling of cross-sectional data, we need an example of where e_i is correlated with a value x_i corresponding to the same ith observation.

Assumptions of the Simple Linear Regression Model Under Random Sampling

RS1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

RS2: The data pairs (y_i, x_i) are statistically independent of all other data pairs and have the same joint distribution $f(y_i, x_i)$. They are independent and identically distributed.

RS3: $E(e_i|x_i) = 0$ for $i = 1, \ldots, N$; x is strictly exogenous.

RS4: The random error has constant conditional variance, $\text{var}(e_i|x_i) = \sigma^2$.

RS5: x_i takes at least two different values.

RS6: $e_i \sim N(0, \sigma^2)$.

Example 4. Suppose that x_i is a measure of the quantity of inputs used in a production process by a randomly chosen firm in an equation designed to explain a firm's production costs. The error term e_i may contain unmeasured features associated with the ability of the firm's managers. It is possible that more able managers are able to use fewer inputs in the production process, so we might expect $\text{cov}(x_i, e_i) < 0$. In this case, strict exogeneity fails. The ith firm's input usage is correlated with unmeasured characteristics of firm managers contained in the ith error, e_i. A firm's input usage is not strictly exogenous, and in econometric terms, it is said to be **endogenous**. Explanatory variables are *endogenous* if they are correlated with the error term.

[13] http://www.bls.gov/nls/home.htm

2.11 Exercises

2.11.1 Problems

2.1 Consider the following five observations. You are to do all the parts of this exercise using only a calculator.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	4				
2	2				
1	3				
–1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum(x_i-\bar{x}) =$	$\sum(x_i-\bar{x})^2 =$	$\sum(y_i-\bar{y}) =$	$\sum(x_i-\bar{x})(y_i-\bar{y}) =$

a. Complete the entries in the table. Put the sums in the last row. What are the sample means $\bar{x}$ and $\bar{y}$?

b. Calculate b_1 and b_2 using (2.7) and (2.8) and state their interpretation.

c. Compute $\sum_{i=1}^{5} x_i^2, \sum_{i=1}^{5} x_i y_i$. Using these numerical values, show that $\sum(x_i-\bar{x})^2 = \sum x_i^2 - N\bar{x}^2$ and $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - N\bar{x}\bar{y}$.

d. Use the least squares estimates from part (b) to compute the fitted values of y, and complete the remainder of the table below. Put the sums in the last row.
Calculate the sample variance of y, $s_y^2 = \sum_{i=1}^{N}(y_i-\bar{y})^2/(N-1)$, the sample variance of x, $s_x^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2/(N-1)$, the sample covariance between x and y, $s_{xy} = \sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})/(N-1)$, the sample correlation between x and y, $r_{xy} = s_{xy}/(s_x s_y)$ and the coefficient of variation of x, $CV_x = 100(s_x/\bar{x})$. What is the median, 50th percentile, of x?

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	4				
2	2				
1	3				
–1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum \hat{y}_i =$	$\sum \hat{e}_i =$	$\sum \hat{e}_i^2 =$	$\sum x_i\hat{e}_i =$

e. On graph paper, plot the data points and sketch the fitted regression line $\hat{y}_i = b_1 + b_2 x_i$.

f. On the sketch in part (e), locate the point of the means $(\bar{x}, \bar{y})$. Does your fitted line pass through that point? If not, go back to the drawing board, literally.

g. Show that for these numerical values $\bar{y} = b_1 + b_2\bar{x}$.

h. Show that for these numerical values $\bar{\hat{y}} = \bar{y}$, where $\bar{\hat{y}} = \sum \hat{y}_i/N$.

i. Compute $\hat{\sigma}^2$.

j. Compute $\widehat{\text{var}}(b_2|\mathbf{x})$ and $\text{se}(b_2)$.

2.2 A household has weekly income of \$2000. The mean weekly expenditure for households with this income is $E(y|x = \$2000) = \mu_{y|x=\$2000} = \$220$, and expenditures exhibit variance $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = \$121$.

a. Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$200 and \$215 on food in a week. Include a sketch with your solution.

b. Find the probability that a household with this income spends more than $250 on food in a week. Include a sketch with your solution.

c. Find the probability in part (a) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 144$.

d. Find the probability in part (b) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 144$.

2.3 Graph the following observations of x and y on graph paper.

TABLE 2.4 Exercise 2.3 Data

x	1	2	3	4	5	6
y	6	4	11	9	13	17

a. Using a ruler, draw a line that fits through the data. Measure the slope and intercept of the line you have drawn.

b. Use formulas (2.7) and (2.8) to compute, using only a hand calculator, the least squares estimates of the slope and the intercept. Plot this line on your graph.

c. Obtain the sample means $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$. Obtain the predicted value of y for $x = \bar{x}$ and plot it on your graph. What do you observe about this predicted value?

d. Using the least squares estimates from (b), compute the least squares residuals $\hat{e}_i$.

e. Find their sum, $\sum \hat{e}_i$, and their sum of squared values, $\sum \hat{e}_i^2$.

f. Calculate $\sum x_i \hat{e}_i$.

2.4 We have defined the simple linear regression model to be $y = \beta_1 + \beta_2 x + e$. Suppose, however, that we knew, for a fact, that $\beta_1 = 0$.

a. What does the linear regression model look like, algebraically, if $\beta_1 = 0$?

b. What does the linear regression model look like, graphically, if $\beta_1 = 0$?

c. If $\beta_1 = 0$, the least squares "sum of squares" function becomes $S(\beta_2) = \sum_{i=1}^{N} (y_i - \beta_2 x_i)^2$. Using the data in Table 2.4 from Exercise 2.3, plot the value of the sum of squares function for enough values of β_2 for you to locate the approximate minimum. What is the significance of the value of β_2 that minimizes $S(\beta_2)$? [*Hint*: Your computations will be simplified if you algebraically expand $S(\beta_2) = \sum_{i=1}^{N} (y_i - \beta_2 x_i)^2$ by squaring the term in parentheses and carrying through the summation operator.]

d. Using calculus, show that the formula for the least squares estimate of β_2 in this model is $b_2 = \sum x_i y_i / \sum x_i^2$. Use this result to compute b_2 and compare this value with the value you obtained geometrically.

e. Using the estimate obtained with the formula in (d), plot the fitted (estimated) regression function. On the graph locate the point $(\bar{x}, \bar{y})$. What do you observe?

f. Using the estimate obtained with the formula in (d), obtain the least squares residuals, $\hat{e}_i = y_i - b_2 x_i$. Find their sum.

g. Calculate $\sum x_i \hat{e}_i$.

2.5 A small business hires a consultant to predict the value of weekly sales of their product if their weekly advertising is increased to $2000 per week. The consultant takes a record of how much the firm spent on advertising per week and the corresponding weekly sales over the past six months. The consultant writes, "Over the past six months the average weekly expenditure on advertising has been $1500 and average weekly sales have been $10,000. Based on the results of a simple linear regression, I predict sales will be $12,000 if $2000 per week is spent on advertising."

a. What is the estimated simple regression used by the consultant to make this prediction?

b. Sketch a graph of the estimated regression line. Locate the average weekly values on the graph.

2.6 A soda vendor at Louisiana State University football games observes that the warmer the temperature at game time the greater the number of sodas that are sold. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 20x$, where y = the number of sodas she sells and x = temperature in degrees Fahrenheit.

a. Interpret the estimated slope and intercept. Do the estimates make sense? Why or why not?

b. On a day when the temperature at game time is forecast to be 80°F, predict how many sodas the vendor will sell.
c. Below what temperature are the predicted sales zero?
d. Sketch a graph of the estimated regression line.

2.7 We have 2008 data on y = income per capita (in thousands of dollars) and x = percentage of the population with a bachelor's degree or more for the 50 U.S. states plus the District of Columbia, a total of $N = 51$ observations. We have results from a simple linear regression of y on x.

a. The estimated error variance is $\hat{\sigma}^2 = 14.24134$. What is the sum of squared least squares residuals?
b. The estimated variance of b_2 is 0.009165. What is the standard error of b_2? What is the value of $\sum(x_i - \bar{x})^2$?
c. The estimated slope is $b_2 = 1.02896$. Interpret this result.
d. Using $\bar{x} = 27.35686$ and $\bar{y} = 39.66886$, calculate the estimate of the intercept.
e. Given the results in (b) and (d), what is $\sum x_i^2$?
f. For the state of Georgia, the value of $y = 34.893$ and $x = 27.5$. Compute the least squares residual, using the information in parts (c) and (d).

2.8 Professor I.M. Mean likes to use averages. When fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i), Professor Mean calculates the sample means (averages) of (y_i, x_i) for the first three and second three observations in the data $\left(\bar{y}_1 = \sum_{i=1}^{3} y_i/3,\ \bar{x}_1 = \sum_{i=1}^{3} x_i/3\right)$ and $\left(\bar{y}_2 = \sum_{i=4}^{6} y_i/3,\ \bar{x}_2 = \sum_{i=4}^{6} x_i/3\right)$. Then Dr. Mean's estimator of the slope is $\hat{\beta}_{2,mean} = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$ and the Dr. Mean intercept estimator is $\hat{\beta}_{1,mean} = \bar{y} - \hat{\beta}_{2,mean}\bar{x}$, where $(\bar{y}, \bar{x})$ are the sample means using all the data. You may use a spreadsheet or other software to carry out tedious calculations.

a. Calculate $\hat{\beta}_{1,mean}$ and $\hat{\beta}_{2,mean}$. Plot the data, and the fitted line $\hat{y}_{i,mean} = \hat{\beta}_{1,mean} + \hat{\beta}_{2,mean} x_i$.
b. Calculate the residuals $\hat{e}_{i,mean} = y_i - \hat{y}_{i,mean} = y_i - \left(\hat{\beta}_{1,mean} + \hat{\beta}_{2,mean} x_i\right)$. Find $\sum_{i=1}^{6} \hat{e}_{i,mean}$, and $\sum_{i=1}^{6} x_i \hat{e}_{i,mean}$.
c. Compare the results in (b) to the corresponding values based on the least squares regression estimates. See Exercise 2.3.
d. Compute $\sum_{i=1}^{6} \hat{e}_{i,mean}^2$. Is this value larger or smaller than the sum of squared least squares residuals in Exercise 2.3(d)?

2.9 Professor I.M. Mean likes to use averages. When fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i), Professor Mean calculates the sample means (averages) of (y_i, x_i) for the first three and second three observations in the data $\left(\bar{y}_1 = \sum_{i=1}^{3} y_i/3,\ \bar{x}_1 = \sum_{i=1}^{3} x_i/3\right)$ and $\left(\bar{y}_2 = \sum_{i=4}^{6} y_i/3,\ \bar{x}_2 = \sum_{i=4}^{6} x_i/3\right)$. Then Dr. Mean's estimator of the slope is $\hat{\beta}_{2,mean} = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$.

a. Assuming assumptions SR1–SR6 hold, show that, conditional on $\mathbf{x} = (x_1, \ldots, x_6)$, Dr. Mean's estimator is unbiased, $E\left(\hat{\beta}_{2,mean}|\mathbf{x}\right) = \beta_2$.
b. Assuming assumptions SR1–SR6 hold, show that $E\left(\hat{\beta}_{2,mean}\right) = \beta_2$.
c. Assuming assumptions SR1–SR6 hold, find the theoretical expression for $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$. Is this variance larger or smaller than the variance of the least squares estimator $\text{var}(b_2|\mathbf{x})$? Explain.

2.10 Consider fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i). Suppose that based on a theoretical argument we **know** that $\beta_2 = 0$.

a. What does the regression model look like, algebraically, if $\beta_2 = 0$?
b. What does the regression model look like, graphically, if $\beta_2 = 0$?
c. If $\beta_2 = 0$ the sum of squares function becomes $S(\beta_1) = \sum_{i=1}^{N} (y_i - \beta_1)^2$. Using the data in Table 2.4, plot the sum of squares function for enough values of β_1 so that you can locate the approximate minimum. What is this value? [*Hint*: Your calculations will be easier if you square the term in parentheses and carry through the summation operator.]
d. Using calculus, show that the formula for the least squares estimate of β_1 in this model is $\hat{\beta}_1 = \left(\sum_{i=1}^{N} y_i\right)/N$.
e. Using the data in Table 2.4 and the result in part (d), compute an estimate of β_1. How does this value compare to the value you found in part (c)?

f. Using the data in Table 2.4, calculate the sum of squared residuals $S\left(\hat{\beta}_1\right) = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_1\right)^2$. Is this sum of squared residuals larger or smaller than the sum of squared residuals $S\left(b_1, b_2\right) = \sum_{i=1}^{N}\left(y_i - b_1 - b_2 x_i\right)^2$ using the least squares estimates? [See Exercise 2.3 (d).]

2.11 Let y = expenditure ($) on food away from home per household member per month in the past quarter and x = monthly household income (in hundreds of dollars) during the past year.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 13.77 + 0.52x$. Interpret the estimated slope and intercept from this relation.

b. Predict the expenditures on food away from home for a household with $2000 a month income.

c. Calculate the elasticity of expenditure on food away from home with respect to income when household income is $2000 per month. [*Hint*: Elasticity must be calculated for a point on the fitted regression.]

d. We estimate the log-linear model to be $\widehat{\ln(y)} = 3.14 + 0.007x$. What is the estimated elasticity of expenditure on food away from home with respect to income, if household income is $2000 per month?

e. For the log-linear model in part (d), calculate $\hat{y} = \exp(3.14 + 0.007x)$ when $x = 20$ and when $x = 30$. Evaluate the slope of the relation between y and x, dy/dx, for each of these $\hat{y}$ values. Based on these calculations for the log-linear model, is expenditure on food away from home increasing with respect to income at an increasing or decreasing rate?

f. When estimating the log-linear model in part (d), the number of observations used in the regression falls to $N = 2005$. How many households in the sample reported no expenditures on food away from home in the past quarter?

2.12 Let y = expenditure ($) on food away from home per household member per month in the past quarter and $x = 1$ if the household includes a member with an advanced degree, a Master's, or Ph.D./Professional degree, and $x = 0$ otherwise.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 44.96 + 30.41x$. Interpret the coefficient of x and the intercept from this relation.

b. What is the per person sample mean of food expenditures away from home for a household including someone with an advanced degree?

c. What is the per person sample mean of food expenditures away from home for a household that does not include someone with an advanced degree?

2.13 Using 2011 data on 141 U.S. public research universities, we examine the relationship between academic cost per student, *ACA* (real total academic cost per student in thousands of dollars) and full-time enrollment *FTESTU* (in thousands of students).

a. The least squares fitted relation is $\widehat{ACA} = 14.656 + 0.266FTESTU$. What is the economic interpretation of the estimated parameters? Why isn't the intercept zero?

b. In 2011 Louisiana State University (LSU) had a full-time student enrollment of 27,950. Using the fitted related in part (a), compute the predicted value of *ACA*.

c. The actual value of *ACA* for LSU that year was 21.403. Calculate the least squares residual for LSU? Does the model overpredict or underpredict *ACA* for LSU?

d. The sample mean (average) full-time enrollment in U.S. public research universities in 2011 was 22,845.77. What was the sample mean of academic cost per student?

2.14 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$, where *WAGE* is hourly wage rate in U.S. 2013 dollars and *EDUC* is years of education, or schooling. The regression model is estimated twice using the least squares estimator, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \quad \underset{(\text{se})}{\widehat{WAGE} = -10.76 + 2.46\,EDUC,} \quad \underset{(2.27)\ (0.16)}{N = 986}$$

$$\text{Rural} \quad \underset{(\text{se})}{\widehat{WAGE} = -4.88 + 1.80\,EDUC,} \quad \underset{(3.29)\ (0.24)}{N = 214}$$

a. Using the estimated rural regression, compute the elasticity of wages with respect to education at the "point of the means." The sample mean of *WAGE* is $19.74.

b. The sample mean of *EDUC* in the urban area is 13.68 years. Using the estimated urban regression, compute the standard error of the elasticity of wages with respect to education at the "point of the means." Assume that the mean values are "givens" and not random.

c. What is the predicted wage for an individual with 12 years of education in each area? With 16 years of education?

2.15 Professor E.Z. Stuff has decided that the least squares estimator is too much trouble. Noting that two points determine a line, Dr. Stuff chooses two points from a sample of size N and draws a line between them, calling the slope of this line the EZ estimator of β_2 in the simple regression model. Algebraically, if the two points are (x_1, y_1) and (x_2, y_2), the EZ estimation rule is

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

Assuming that all the assumptions of the simple regression model hold:

a. Show that b_{EZ} is a "linear" estimator.

b. Show that b_{EZ} is an unbiased estimator.

c. Find the conditional variance of b_{EZ}.

d. Find the conditional probability distribution of b_{EZ}.

e. Convince Professor Stuff that the EZ estimator is not as good as the least squares estimator. No proof is required here.

2.11.2 Computer Exercises

2.16 The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally, the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk-free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of *beta* less than one indicate that the stock is "defensive" since its variation is less than the market's. A *beta* greater than one indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e_j$$

a. Explain why the econometric model above is a simple regression model like those discussed in this chapter.

b. In the data file *capm5* are data on the monthly returns of six firms (GE, IBM, Ford, Microsoft, Disney, and Exxon-Mobil), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk-free asset (*RISKFREE*). The 180 observations cover January 1998 to December 2012. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?

c. Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.

d. Estimate the model for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the *beta* values change much?

2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

2.18 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Create histograms for *PRICE* and ln(*PRICE*). Are the distributions skewed or symmetrical?

b. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Interpret the OLS estimates, $\hat{\gamma}_1$ and $\hat{\gamma}_2$. Graph the fitted *PRICE*, $\widehat{PRICE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT\right)$, against *SQFT*, and sketch the tangent line to the curve for a house with 2000 square feet of living area. What is the slope of the tangent line?

c. Compute the least squares residuals from the model in (b) and plot them against *SQFT*. Do any of our assumptions appear violated?

d. Calculate summary statistics for *PRICE* and *SQFT* for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). What differences and/or similarities do you observe?

e. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$ for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). Interpret the estimated coefficient of *SQFT* in each sample's regression.

f. Are the regression results in part (b) valid if the differences you observe in part (e) are substantial? Think in particular about whether SR1 is satisfied.

2.19 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Plot house selling price *SPRICE* against house living area for all houses in the sample.

b. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ for all the houses in the sample. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ for all the houses in the sample. What is the marginal effect of an additional 100 square feet of living area for a home with 1500 square feet of living area.

d. In the same graph, plot the fitted lines from the linear and quadratic models. Which seems to fit the data better? Compare the sum of squared residuals (*SSE*) for the two models. Which is smaller?

e. If the quadratic model is in fact "true," what can we say about the results and interpretations we obtain for the linear relationship in part (b)?

2.20 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: The data file *stockton5* includes 2610 observations.]. Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Finally, estimate the regression using data on both large and small lots. Interpret the estimates. How do the estimates compare?

b. Estimate the regression model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?

c. Estimate a linear regression $SPRICE = \eta_1 + \eta_2 LGELOT + e$ with dependent variable *SPRICE* and independent variable the indicator *LGELOT*, which identifies houses on larger lots. Interpret these results.

d. If the estimates in part (a) and/or part (b) differ substantially for the large lot and small lot subsamples, will assumption SR1 be satisfied in the model that pools all the observations together? If not, why not? Do the results in (c) offer any information about the potential validity of SR1?

2.21 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the linear model $SPRICE = \delta_1 + \delta_2 AGE + e$. Interpret the estimated coefficients. Predict the selling price of a house that is 30 years old.

b. Using the results in part (a), plot house selling price against *AGE* and show the fitted regression line. Based on the plot, does the model fit the data well? Explain.

c. Estimate the log-linear model $\ln(SPRICE) = \theta_1 + \theta_2 AGE + e$. Interpret the estimated slope coefficient.

d. Using the results in part (c), compute $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the OLS estimates. Plot $\widehat{SPRICE}$ against *AGE* (connecting the dots) and *SPRICE* vs. *AGE* in the same graph.

e. Predict the selling price of a house that is 30 years old using $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$.

f. Based on the plots and visual fit of the estimated regression lines, which of the two models in (a) or (c) would you prefer? Explain. For each model calculate $\sum_{i=1}^{1200}\left(SPRICE - \widehat{SPRICE}\right)^2$. Is this at all useful in making a comparison between the models? If so, how?

2.22 A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star5_small*. [Note: The data file *star5* contains more observations and variables.]

a. Using children who are in either a regular-sized class or a small class, estimate the regression model explaining students' combined aptitude scores as a function of class size, $TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + e_i$. Interpret the estimates. Based on this regression result, what do you conclude about the effect of class size on learning?

b. Repeat part (a) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

c. Using children who are in either a regular-sized class or a regular-sized class with a teacher aide, estimate the regression model explaining students' combined aptitude scores as a function of the presence of a teacher aide, $TOTALSCORE = \gamma_1 + \gamma_2 AIDE + e$. Interpret the estimates. Based on this regression result, what do you conclude about the effect on learning of adding a teacher aide to the classroom?

d. Repeat part (c) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

2.23 Professor Ray C. Fair has for a number of years built and updated models that explain and predict the U.S. presidential elections. Visit his website at https://fairmodel.econ.yale.edu/vote2016/index2.htm. See in particular his paper entitled "Presidential and Congressional Vote-Share Equations: November 2010 Update." The basic premise of the model is that the Democratic Party's share of the two-party [Democratic and Republican] popular vote is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for reelection. Fair's data, 26 observations for the election years from 1916 to 2016, are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic Party. Consider the effect of economic growth on *VOTE*. If Democrats are the incumbent party (*INCUMB* = 1) then economic growth, the growth rate in real per capita GDP in the first three quarters of the election year (annual rate), should enhance their chances of winning. On the other hand, if the Republicans are the incumbent party (*INCUMB* = −1), growth will diminish the Democrats' chances of winning. Consequently, we define the explanatory variable *GROWTH* = *INCUMB* × growth rate.

a. Using the data for 1916–2012, plot a scatter diagram of *VOTE* against *GROWTH*. Does there appear to be a positive association?

b. Estimate the regression $VOTE = \beta_1 + \beta_2 GROWTH + e$ by least squares using the data from 1916 to 2012. Report and discuss the estimation result. Plot the fitted line on the scatter diagram from (a).

c. Using the model estimated in (b), predict the 2016 value of *VOTE* based on the actual 2016 value for *GROWTH*. How does the predicted vote for 2016 compare to the actual result?

d. Economy wide inflation may spell doom for the incumbent party in an election. The variable $INFLAT = INCUMB \times$ inflation rate, where the inflation rate is the growth in prices over the first 15 quarters of an administration. Using the data from 1916 to 2012, plot *VOTE* against *INFLAT*.

e. Using the data from 1916 to 2012, report and discuss the estimation results for the model $VOTE = \alpha_1 + \alpha_2 INFLAT + e$.

f. Using the model estimated in (e), predict the 2016 value of *VOTE* based on the actual 2012 value for *INFLAT*. How does the predicted vote for 2016 compare to the actual result?

2.24 Using data on the "Ashcan School"[14] we have an opportunity to study the market for art. What factors determine the value of a work of art? Use the data in *ashcan_small*. [Note: The file *ashcan* contains more variables.] For this exercise, use data only on works that sold ($SOLD = 1$).

a. Using data on works that sold, construct a histogram for *RHAMMER* and compute summary statistics. What are the mean and median prices for the artwork sold? What are the 25th and 75th percentiles?

b. Using data on works that sold, construct a histogram for ln(*RHAMMER*). Describe the shape of this histogram as compared to that in part (a).

c. Plot ln(*RHAMMER*) against the age of the painting at the time of its sale, $YEARS\_OLD = DATE\_AUCTN - CREATION$. Include in the plot the least squares fitted line. What patterns do you observe?

d. Use data on works that sold, estimate the regression $\ln(RHAMMER) = \beta_1 + \beta_2 YEAR\_SOLD + e$. Interpret the estimated coefficient of *YEARS_OLD*.

e. *DREC* is an indicator variable equaling 1 if the work was sold during a recession. Using data on works that sold, estimate the regression $\ln(RHAMMER) = \alpha_1 + \alpha_2 DREC + e$. Interpret the estimated coefficient of *DREC*.

2.25 Consumer expenditure data from 2013 are contained in the file *cex5_small*. [Note: *cex5* is a larger version with more observations and variables.] Data are on three-person households consisting of a husband and wife, plus one other member, with incomes between \$1000 per month to \$20,000 per month. *FOODAWAY* is past quarter's food away from home expenditure per month per person, in dollars, and *INCOME* is household monthly income during past year, in \$100 units.

a. Construct a histogram of *FOODAWAY* and its summary statistics. What are the mean and median values? What are the 25th and 75th percentiles?

b. What are the mean and median values of *FOODAWAY* for households including a member with an advanced degree? With a college degree member? With no advanced or college degree member?

c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

e. Plot ln(*FOODAWAY*) against *INCOME*, and include the fitted line from part (d).

f. Calculate the least squares residuals from the estimation in part (d). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

2.26 Consumer expenditure data from 2013 are contained in the file *cex5_small*. [Note: *cex5* is a larger version with more observations and variables.] Data are on three-person households consisting of a husband and wife, plus one other member, with incomes between \$1000 per month to \$20,000 per month. *FOODAWAY* is past quarter's food away from home expenditure per month per person, in dollars, and *INCOME* is household monthly income during past year, in \$100 units.

a. Estimate the linear regression $FOODAWAY = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

b. Calculate the least squares residuals from the estimation in part (b). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

c. Estimate the linear regression $FOODAWAY = \alpha_1 + \alpha_2 ADVANCED + e$. Interpret the estimated coefficient of *ADVANCED*.

d. What are the sample means of *FOODAWAY* for households including a member with an advanced degree? With no advanced degree member? How do these values relate to the regression in part (c)?

[14] Robert B. Ekelund, Jr., John D. Jackson, and Robert D. Tollison "Are Art Auction Estimates Biased" published in *Southern Economic Journal*, 80(2), 2013, 454–465; also http://en.wikipedia.org/wiki/Ashcan_School

2.27 The owners of a motel discovered that a defective product was used in its construction. It took seven months to correct the defects during which 14 rooms in the 100-unit motel were taken out of service for one month at a time. For this exercise use the data file *motel*.

a. Graph $y = MOTEL\_PCT$, percentage motel occupancy, against $x = 100RELPRICE$, which is the percentage of the competitor's price per room charged by the motel in question. Describe the relationship between the variables based on the graph. Is there a positive association, an inverse association, or no association?

b. Consider the linear regression $MOTEL\_PCT_t = \beta_1 + \beta_2 100RELPRICE_t + e_t$. What sign do you predict for the slope coefficient? Why? Does the sign of the estimated slope agree with your expectation?

c. Calculate the least squares residuals from the regression in (b). Plot the residuals against $TIME = 1, \ldots, 25$ (month 1 = March 2003, … , month 25 = March 2005). On the graph indicate residuals when $TIME = 17, 18, \ldots, 23$. These are the months of repair. Does the model overpredict or underpredict the motel's occupancy rates for those months?

d. Estimate the linear regression $MOTEL\_PCT_t = \alpha_1 + \alpha_2 REPAIR_t + e_t$, where $REPAIR_t = 1$ for months when repairs were occurring and $REPAIR_t = 0$ otherwise. What was the motel's mean occupancy rate when there were no repairs being made? What was the motel's mean occupancy rate when repairs were being made?

2.28 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

a. Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics.

b. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results.

c. Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?

d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

e. Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).

f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

2.29 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version with more observations and variables.]

a. Create the variable $LWAGE = \ln(WAGE)$. Construct a histogram and calculate detailed summary statistics. Does the histogram appear bell shaped and normally distributed? A normal distribution is symmetrical with no skewness, *skewness* = 0. The tails of the normal distribution have a certain "thickness." A measure of the tail thickness is *kurtosis*, discussed in Appendix C.4.2. For a normal distribution, the *kurtosis* = 3, discussed in Appendix C.7.4. How close are the measures of *skewness* and *kurtosis* for *LWAGE* to 0 and 3, respectively?

b. Obtain the OLS estimates from the log-linear regression model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$ and interpret the estimated value of β_2.

c. Obtain the predicted wage, $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$, for a person with 12 years of education and for a person with 16 years of education.

d. What is the marginal effect of additional education for a person with 12 years of education and for a person with 16 years of education? [*Hint*: This is the slope of the fitted model at those two points.]

e. Plot the fitted values $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$ versus *EDUC* in a graph. Also include in the graph the fitted linear relationship. Based on the graph, which model seems to fit the data better, the linear or log-linear model?

f. Using the fitted values from the log-linear model, compute $\sum\left(WAGE - \widehat{WAGE}\right)^2$. Compare this value to the sum of squared residuals from the estimated linear relationship. Using this as a basis of comparison, which model fits the data better?

2.30 In this exercise, we consider the amounts that are borrowed for single family home purchases in Las Vegas, Nevada, during 2010. Use the data file *vegas5_small* for this exercise.

- **a.** Compute summary statistics for *AMOUNT*, *FICO, RATE*, and *TERM30*. What is the sample average amount borrowed? What *FICO* score corresponds to the 90th percentile? What is the median interest rate paid, and what percent of the mortgages were for 30-year terms?
- **b.** Construct histograms for *AMOUNT*, ln(*AMOUNT*), *FICO*, and *RATE*. Are the empirical distributions symmetrical? Do they have one peak (unimodal) or two peaks (bimodal)?
- **c.** Estimate regressions for dependent variables *AMOUNT* and ln(*AMOUNT*) against the independent variable *FICO*. For each regression, interpret the coefficient of *FICO*.
- **d.** Estimate regressions for dependent variables *AMOUNT* and ln(*AMOUNT*) against the independent variable *RATE*. For each regression, interpret the coefficient of *RATE*.
- **e.** Estimate a regression with dependent variable *AMOUNT* and explanatory variable *TERM30*. Obtain the summary statistics for *AMOUNT* for transactions with 30-year loans and for those transactions when the term was not 30 years. Explain the regression results in terms of the summary statistics you have calculated.

Appendix 2A Derivation of the Least Squares Estimates

Given the sample observations on y and x, we want to find values for the unknown parameters β_1 and β_2 that minimize the "sum of squares" function

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2 \tag{2A.1}$$

Since the points (y_i, x_i) have been observed, the sum of squares function S depends only on the unknown parameters β_1 and β_2. This function, which is a quadratic in terms of the unknown parameters β_1 and β_2, is a "bowl-shaped surface" like the one depicted in Figure 2A.1.

Our task is to find, out of all the possible values β_1 and β_2, the point (b_1, b_2) at which the sum of squares function S is a minimum. This minimization problem is a common one in calculus, and the minimizing point is at the "bottom of the bowl."

Those of you familiar with calculus and "partial differentiation" can verify that the partial derivatives of S with respect to β_1 and β_2 are

$$\frac{\partial S}{\partial \beta_1} = 2N\beta_1 - 2\sum y_i + 2\left(\sum x_i\right)\beta_2$$

$$\frac{\partial S}{\partial \beta_2} = 2\left(\sum x_i^2\right)\beta_2 - 2\sum x_i y_i + 2\left(\sum x_i\right)\beta_1 \tag{2A.2}$$

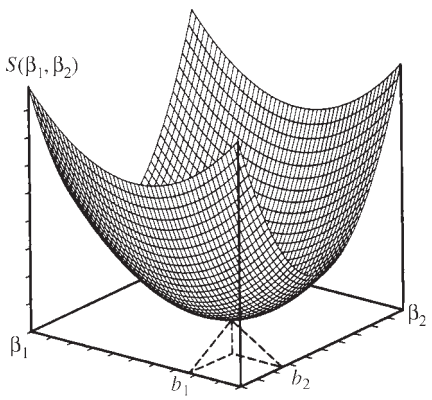


FIGURE 2A.1 **The sum of squares function and the minimizing values b_1 and b_2.**

These derivatives are equations of the slope of the bowl-like surface in the directions of the axes. Intuitively, the "bottom of the bowl" occurs where the slope of the bowl, in the direction of each axis, $\partial S/\partial\beta_1$ and $\partial S/\partial\beta_2$, is zero.

Algebraically, to obtain the point (b_1, b_2), we set (2A.2) to zero and replace β_1 and β_2 by b_1 and b_2, respectively, to obtain

$$2\left[\sum y_i - Nb_1 - \left(\sum x_i\right)b_2\right] = 0$$

$$2\left[\sum x_i y_i - \left(\sum x_i\right)b_1 - \left(\sum x_i^2\right)b_2\right] = 0$$

Simplifying these gives equations usually known as the **normal equations**:

$$Nb_1 + \left(\sum x_i\right)b_2 = \sum y_i \tag{2A.3}$$

$$\left(\sum x_i\right)b_1 + \left(\sum x_i^2\right)b_2 = \sum x_i y_i \tag{2A.4}$$

These two equations have two unknowns b_1 and b_2. We can find the least squares estimates by solving these two linear equations for b_1 and b_2. To solve for b_2, multiply (2A.3) by $\sum x_i$, multiply (2A.4) by N, then subtract the first equation from the second, and then isolate b_2 on the left-hand side.

$$b_2 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{2A.5}$$

This formula for b_2 is in terms of data sums, cross-products, and squares. The deviation from the mean form of the estimator is derived in Appendix 2B.

To solve for b_1, given b_2, divide both sides of (2A.3) by N and rearrange.

Appendix 2B Deviation from the Mean Form of b_2

The first step in the conversion of the formula for b_2 into (2.7) is to use some tricks involving summation signs. The first useful fact is that

$$\begin{aligned}\sum\left(x_i - \bar{x}\right)^2 &= \sum x_i^2 - 2\bar{x}\sum x_i + N\bar{x}^2 = \sum x_i^2 - 2\bar{x}\left(N\tfrac{1}{N}\sum x_i\right) + N\bar{x}^2 \\ &= \sum x_i^2 - 2N\bar{x}^2 + N\bar{x}^2 = \sum x_i^2 - N\bar{x}^2\end{aligned} \tag{2B.1}$$

Should you ever have to calculate $\sum\left(x_i - \bar{x}\right)^2$, using the shortcut formula $\sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - N\bar{x}^2$ is usually much easier. Then

$$\sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - N\bar{x}^2 = \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N} \tag{2B.2}$$

To obtain this result, we have used the fact that $\bar{x} = \sum x_i/N$, so $\sum x_i = N\bar{x}$.

The second useful fact is similar to the first, and it is

$$\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - N\bar{x}\bar{y} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{N} \tag{2B.3}$$

This result is proven in a similar manner.

If the numerator and denominator of b_2 in (2A.5) are divided by N, then using (2B.1)–(2B.3), we can rewrite b_2 in *deviation from the mean form* as

$$b_2 = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}$$

This formula for b_2 is one that you should remember, as we will use it time and time again in the next few chapters.

Appendix 2C b_2 Is a Linear Estimator

In order to derive (2.10), we make a further simplification using another property of sums. The sum of any variable about its average is zero; that is,

$$\sum(x_i - \bar{x}) = 0$$

Then, the formula for b_2 becomes

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$
$$= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum\left[\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] y_i = \sum w_i y_i$$

where w_i is given in (2.11).

Appendix 2D Derivation of Theoretical Expression for b_2

To obtain (2.12) replace y_i in (2.10) by $y_i = \beta_1 + \beta_2 x_i + e_i$ and simplify:

$$b_2 = \sum w_i y_i = \sum w_i(\beta_1 + \beta_2 x_i + e_i)$$
$$= \beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i e_i$$
$$= \beta_2 + \sum w_i e_i$$

We used two more summation tricks to simplify this. First, $\sum w_i = 0$; this eliminates the term $\beta_1 \sum w_i$. Secondly, $\sum w_i x_i = 1$, so $\beta_2 \sum w_i x_i = \beta_2$, and (2.10) simplifies to (2.12).

The term $\sum w_i = 0$ because

$$\sum w_i = \sum\left[\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] = \frac{1}{\sum(x_i - \bar{x})^2}\sum(x_i - \bar{x}) = 0$$

where in the last step we used the fact that $\sum(x_i - \bar{x}) = 0$.

To show that $\sum w_i x_i = 1$ we again use $\sum(x_i - \bar{x}) = 0$. Another expression for $\sum(x_i - \bar{x})^2$ is

$$\sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})(x_i - \bar{x})$$
$$= \sum(x_i - \bar{x})x_i - \bar{x}\sum(x_i - \bar{x})$$
$$= \sum(x_i - \bar{x})x_i$$

Consequently,

$$\sum w_i x_i = \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})x_i} = 1$$

Appendix 2E Deriving the Conditional Variance of b_2

The starting point is equation (2.12), $b_2 = \beta_2 + \sum w_i e_i$. The least squares estimator is a random variable whose conditional variance is defined to be

$$\text{var}(b_2|\mathbf{x}) = E\left\{\left[b_2 - E(b_2|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\}$$

Substituting in (2.12) and using the conditional unbiasedness of the least squares estimator, $E(b_2|\mathbf{x}) = \beta_2$, we have

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= E\left\{ \left[\beta_2 + \sum w_i e_i - \beta_2\right]^2 \middle| \mathbf{x} \right\} \\
&= E\left\{ \left[\sum w_i e_i\right]^2 \middle| \mathbf{x} \right\} \\
&= E\left\{ \left[\sum w_i^2 e_i^2 + \sum\sum_{i \neq j} w_i w_j e_i e_j \right] \middle| \mathbf{x} \right\} \qquad \text{[square of bracketed term]} \\
&= E\left\{ \left[\sum w_i^2 e_i^2\right] \middle| \mathbf{x} \right\} + E\left\{ \left[\sum\sum_{i \neq j} w_i w_j e_i e_j \right] \middle| \mathbf{x} \right\} \\
&= \sum w_i^2 E(e_i^2|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j E(e_i e_j|\mathbf{x}) \qquad \text{[because } w_i \text{ not random given } \mathbf{x}\text{]} \\
&= \sigma^2 \sum w_i^2 \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

The next to last line is obtained by using two assumptions: First,

$$\sigma^2 = \text{var}(e_i|\mathbf{x}) = E\left\{ \left[e_i - E(e_i|\mathbf{x})\right]^2 \middle| \mathbf{x} \right\} = E\left[\left(e_i - 0\right)^2 \middle| \mathbf{x} \right] = E(e_i^2|\mathbf{x})$$

Second, $\text{cov}(e_i, e_j|\mathbf{x}) = E\left\{ \left[e_i - E(e_i|\mathbf{x})\right]\left[e_j - E(e_j|\mathbf{x})\right] \middle| \mathbf{x} \right\} = E(e_i e_j|\mathbf{x}) = 0$. Then, the very last step uses the fact that

$$\sum w_i^2 = \sum \left[\frac{(x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} \right] = \frac{\sum (x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} = \frac{1}{\sum (x_i - \bar{x})^2}$$

Alternatively, we can employ the rule for finding the variance of a sum. If X and Y are random variables, and a and b are constants, then

$$\text{var}(aX + bY) = a^2 \text{ var}(X) + b^2 \text{ var}(Y) + 2ab\text{cov}(X, Y)$$

Appendix B.4 reviews all the basic properties of random variables. In the second line below we use this rule extended to more than two random variables. Then,

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= \text{var}\left[\left(\beta_2 + \sum w_i e_i\right) \middle| \mathbf{x} \right] && \text{[since } \beta_2 \text{ is a constant]} \\
&= \sum w_i^2 \text{var}(e_i|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j \text{cov}(e_i, e_j|\mathbf{x}) && \text{[generalizing the variance rule]} \\
&= \sum w_i^2 \text{var}(e_i|\mathbf{x}) && \text{[using } \text{cov}(e_i, e_j|\mathbf{x}) = 0\text{]} \\
&= \sigma^2 \sum w_i^2 && \text{[using } \text{var}(e_i|\mathbf{x}) = \sigma^2\text{]} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

Carefully note that the derivation of the variance expression for b_2 depends on assumptions SR3 and SR4. If the $\text{cov}(e_i, e_j|\mathbf{x}) \neq 0$, then we cannot drop out all those terms in the double summation. If $\text{var}(e_i|\mathbf{x}) \neq \sigma^2$ for all observations, then σ^2 cannot be factored out of the summation. If either of these assumptions fails to hold, then the conditional variance $\text{var}(b_2|\mathbf{x})$ is *something else*, and is not given by (2.15). The same is true for the conditional variance of b_1 and the conditional covariance between b_1 and b_2.

Appendix 2F Proof of the Gauss–Markov Theorem

We will prove the Gauss–Markov theorem for the least squares estimator b_2 of β_2. Our goal is to show that in the class of linear and unbiased estimators the estimator b_2 has the smallest variance. Let $b_2^* = \sum k_i y_i$ (where k_i are constants) be any other linear estimator of β_2. To make comparison to the least squares estimator b_2 easier, suppose that $k_i = w_i + c_i$, where c_i is another constant and w_i is given in (2.11). While this is tricky, it is legal, since for any k_i that someone might choose we can find c_i. Into this new estimator, substitute y_i and simplify, using the properties of w_i in Appendix 2D.

$$\begin{aligned} b_2^* &= \sum k_i y_i = \sum (w_i + c_i) y_i = \sum (w_i + c_i)(\beta_1 + \beta_2 x_i + e_i) \\ &= \sum (w_i + c_i)\beta_1 + \sum (w_i + c_i)\beta_2 x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum w_i + \beta_1 \sum c_i + \beta_2 \sum w_i x_i + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \end{aligned} \tag{2F.1}$$

since $\sum w_i = 0$ and $\sum w_i x_i = 1$.

Take the mathematical expectation of the last line in (2F.1), using the properties of expectation and the assumption that $E(e_i|\mathbf{x}) = 0$:

$$\begin{aligned} E(b_2^*|\mathbf{x}) &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) E(e_i|\mathbf{x}) \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i \end{aligned} \tag{2F.2}$$

In order for the linear estimator $b_2^* = \sum k_i y_i$ to be unbiased, it must be true that

$$\sum c_i = 0 \quad \text{and} \quad \sum c_i x_i = 0 \tag{2F.3}$$

These conditions must hold in order for $b_2^* = \sum k_i y_i$ to be in the class of *linear* and *unbiased estimators*. So we will assume that conditions (2F.3) hold and use them to simplify expression (2F.1):

$$b_2^* = \sum k_i y_i = \beta_2 + \sum (w_i + c_i) e_i \tag{2F.4}$$

We can now find the variance of the linear unbiased estimator b_2^* following the steps in Appendix 2E and using the additional fact that

$$\sum c_i w_i = \sum \left[\frac{c_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sum (x_i - \bar{x})^2} \sum c_i x_i - \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \sum c_i = 0$$

Use the properties of variance to obtain

$$\begin{aligned} \text{var}(b_2^*|\mathbf{x}) &= \text{var}\left\{ \left[\beta_2 + \sum (w_i + c_i) e_i \right] |\mathbf{x} \right\} = \sum (w_i + c_i)^2 \text{var}(e_i|\mathbf{x}) \\ &= \sigma^2 \sum (w_i + c_i)^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2 \\ &= \text{var}(b_2|\mathbf{x}) + \sigma^2 \sum c_i^2 \\ &\geq \text{var}(b_2|\mathbf{x}) \end{aligned}$$

The last line follows since $\sum c_i^2 \geq 0$ and establishes that for the family of linear and unbiased estimators b_2^*, each of the alternative estimators has variance that is greater than or equal to that of the least squares estimator b_2. The *only* time that $\text{var}(b_2^*) = \text{var}(b_2)$ is when all the $c_i = 0$, in which case $b_2^* = b_2$. Thus, there is no *other linear and unbiased estimator of* β_2 that is better than b_2, which proves the Gauss–Markov theorem.

Appendix 2G Proofs of Results Introduced in Section 2.10

2G.1 The Implications of Strict Exogeneity

First, if x is strictly exogenous, then the unconditional expected value of the error term e_i is zero. To show this, we use the law of iterated expectations

$$E(e_i) = E_{x_j}\left[E(e_i|x_j)\right] = E_{x_j}(0) = 0$$

Second, the covariance between X and Y can be calculated as $\text{cov}(X, Y) = E_X\left[(X - \mu_x)\,E(Y|X)\right]$, as discussed in Probability Primer Section P.6.5. Using this result, we obtain

$$\text{cov}(x_j, e_i) = E_{x_j}\left\{\left[x_j - E(x_j)\right]E(e_i|x_j)\right\} = E_{x_j}\left\{\left[x_j - E(x_j)\right]0\right\} = 0$$

If x is strictly exogenous, then the covariance between x_j and e_i is zero for all values of i and j. Recall that zero covariance means "no linear association" but not statistical independence. Thus, strict exogeneity rules out any covariance, any linear association, between any x_j and any e_i. The covariance between x_j and e_i can be rewritten as a simpler expectation using the facts that $E(e_i) = 0$ and $E(x_j)$ is not random

$$\begin{aligned}\text{cov}(x_j, e_i) &= E\left\{\left[x_j - E(x_j)\right]\left[e_i - E(e_i)\right]\right\} = E\left\{\left[x_j - E(x_j)\right]e_i\right\} = E(x_j e_i) - E\left[E(x_j)\,e_i\right] \\ &= E(x_j e_i) - E(x_j)\,E(e_i) = E(x_j e_i)\end{aligned}$$

Strict exogeneity implies $E(x_j e_i) = 0$ for all x_j and e_i.

Using the covariance decomposition we can show yet more. Let $g(x_j)$ be a function of x_j. Then

$$\begin{aligned}\text{cov}\left[g(x_j), e_i\right] &= E_{x_j}\left\{\left[g(x_j) - E\left(g(x_j)\right)\right]E(e_i|x_j)\right\} = E_{x_j}\left\{\left[g(x_j) - E\left(g(x_j)\right)\right]0\right\} = 0 \\ &= E\left[g(x_j)e_i\right]\end{aligned}$$

If x is strictly exogenous, then the covariance between a function of x_j [like x_j^2 or $\ln(x_j)$] and e_i is zero for all values of i and j. Thus, strict exogeneity rules out any covariance, any linear association, between a function of x_j and any e_i.

2G.2 The Random and Independent *x* Case

In Section 2.10.1 we considered the case in which x-values are random but statistically independent of the random error e. In this appendix, we show the algebra behind our conclusions. Consider b_2 the least squares estimator of the slope parameter β_2. b_2 is a linear estimator and as shown in (2.10) $b_2 = \sum_{i=1}^N w_i y_i$, where $w_i = (x_i - \bar{x})/\sum_{i=1}^N (x_i - \bar{x})^2$. Notice that $w_i = g(x_1, \ldots, x_N)$ is a function of all the random x_i values and it is random. For notational ease, let $\mathbf{x}$ represent $x_1, \ldots, x_N$ so $w_i = g(x_1, \ldots, x_N) = g(\mathbf{x})$. Because IRX5 makes clear that x_i is random and is statistically independent of the random error e_i for all values of i and j, then $w_i = g(\mathbf{x})$ is statistically independent of each random error e_i. Substituting $y_i = \beta_1 + \beta_2 x_i + e_i$, we obtain $b_2 = \beta_2 + \sum w_i e_i$ and, using the fact $E(w_i e_i) = E(w_i)E(e_i)$ because of independence, we have

$$E(b_2) = \beta_2 + \sum E(w_i e_i) = \beta_2 + \sum E(w_i)E(e_i) = \beta_2 + \sum E(w_i)\,0 = \beta_2$$

In the case in which x is random but statistically independent of the error terms, the least squares estimator is unconditionally unbiased.

The derivation of the variance of the least squares estimator changes in a similar way:

$$\begin{aligned}\operatorname{var}(b_2) &= E\left[(b_2-\beta_2)^2\right] = E\left[(\beta_2+\sum w_i e_i-\beta_2)^2\right] = E\left[\left(\sum w_i e_i\right)^2\right] \\ &= E\left(\sum w_i^2 e_i^2 + \sum\sum_{i\neq j} w_i w_j e_i e_j\right) \\ &= \sum E(w_i^2)E(e_i^2) + \sum\sum_{i\neq j} E(w_i w_j)\,E(e_i e_j) \\ &= \sigma^2\sum E(w_i^2) = \sigma^2 E\left(\sum w_i^2\right) = \sigma^2 E\left[\frac{1}{\sum(x_i-\bar{x})^2}\right]\end{aligned}$$

In the third line we used the statistical independence of w_i and each random error e_i twice. In the fourth line we used the fact that the expected value of a sum is the sum of the expected values, and finally that $\sum w_i^2$ is known, as shown in Appendix 2E.

The usual estimator of the error variance is $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-2)$ and conditional on $\mathbf{x}$ this estimator is unbiased, $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$. The proof is messy and not shown. This is a conditional expectation saying *given* $x_1, \ldots, x_N$ the estimator $\hat{\sigma}^2$ is unbiased. Now we use the law of iterated expectations from the Probability Primer Section P.6.3:

$$E\left(\hat{\sigma}^2\right) = E_{\mathbf{x}}\left[E\left(\hat{\sigma}^2|\mathbf{x}\right)\right] = E_{\mathbf{x}}\left[\sigma^2\right] = \sigma^2$$

where $E_{\mathbf{x}}(\,)$ means the expected value treating $\mathbf{x}$ as random. Because the conditional expectation $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$ is a constant that does not depend on $\mathbf{x}$, its expectation treating $\mathbf{x}$ as random is also a constant, σ^2. So, in the case in which x is random and independent of the error, $\hat{\sigma}^2$ is conditionally *and* unconditionally unbiased.

The variance of the least squares estimator is

$$\operatorname{var}(b_2) = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum(x_i-\bar{x})^2}\right]$$

The usual variance estimator from (2.21) is

$$\widehat{\operatorname{var}}(b_2|\mathbf{x}) = \hat{\sigma}^2\frac{1}{\sum(x_i-\bar{x})^2}$$

It is an unbiased estimator of $\operatorname{var}(b_2)$ conditional on $\mathbf{x}$. Using the law of iterated expectations, we have

$$E_{\mathbf{x}}\left\{E\left[\widehat{\operatorname{var}}(b_2|\mathbf{x})\right]\right\} = E_{\mathbf{x}}\left\{\sigma^2\frac{1}{\sum(x_i-\bar{x})^2}\,\middle|\,\mathbf{x}\right\} = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum(x_i-\bar{x})^2}\right] = \operatorname{var}(b_2)$$

Thus, the usual estimator of $\operatorname{var}(b_2)$ is unbiased.

What about the Gauss–Markov theorem? It says, for fixed x, or *given* $\mathbf{x}$, $\operatorname{var}(b_2|\mathbf{x})$, is less than the variance $\operatorname{var}(b_2^*|\mathbf{x})$ of any other linear and unbiased estimator b_2^*. That is,

$$\operatorname{var}(b_2|\mathbf{x}) < \operatorname{var}(b_2^*|\mathbf{x})$$

Using the variance decomposition $\operatorname{var}(b_2) = \operatorname{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] + E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right] = E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right]$ because $\operatorname{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] = \operatorname{var}_{\mathbf{x}}(\beta_2) = 0$. Similarly, $\operatorname{var}(b_2^*) = E_{\mathbf{x}}\left[\operatorname{var}(b_2^*|\mathbf{x})\right]$. Then

$$\operatorname{var}(b_2) = E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right] < \operatorname{var}(b_2^*) = E_{\mathbf{x}}\left[\operatorname{var}(b_2^*|\mathbf{x})\right]$$

The logic of the argument is that if $\operatorname{var}(b_2|\mathbf{x})$ is less than the variance of any other estimator $\operatorname{var}(b_2^*|\mathbf{x})$ for any given $\mathbf{x}$, it must also be true for all $\mathbf{x}$, and will remain true if we average over all possible $\mathbf{x}$, by taking the expected value treating $\mathbf{x}$ as random, $E_{\mathbf{x}}(\,)$.

Finally, what about normality? If IRX6 holds, $e_i \sim N(0, \sigma^2)$, then what is the probability distribution of the least squares estimator? We have used the fact that $b_2 = \beta_2 + \sum w_i e_i$. If w_i is constant, then we can assert that the least squares estimator has a normal distribution because linear combinations of normal random variables are normal. However, in the random-x case, even though x is independent of e, the distributions of $w_i e_i$ are not normal. The function $w_i = g(x_1, \ldots, x_N)$ has an unknown probability distribution and its product with the normally distributed e_i results in an unknown distribution. What we can say is that $b_2|\mathbf{x}$ is normal, since conditioning on $x_1, \ldots, x_N$ means that they are treated as given, or fixed.

2G.3 The Random and Strictly Exogenous x Case

In Section 2.10.2 we examine the consequences of an assumption that is weaker than the statistical independence of x and e. There we assert that even with the weaker assumption called "strict exogeneity" the properties of the least squares estimator are unchanged, and here we give the proof. The least squares estimator of the slope parameter, b_2, is a linear estimator and as shown in (2.10) $b_2 = \sum_{i=1}^{N} w_i y_i$, where $w_i = (x_i - \bar{x}) / \sum_{i=1}^{N} (x_i - \bar{x})^2$. Notice that $w_i = g(x_1, \ldots, x_N)$ is a function of all the random x_i values and it is random. Substituting $y_i = \beta_1 + \beta_2 x_i + e_i$, we obtain $b_2 = \beta_2 + \sum w_i e_i$. The strict exogeneity assumption says $E(e_i|x_j) = 0$ for all values of i and j, or equivalently, $E(e_i|\mathbf{x}) = 0$. Using the law of iterated expectations, we show that b_2 is a conditionally unbiased estimator. First, find the conditional expectation of b_2 given $\mathbf{x}$,

$$E(b_2|\mathbf{x}) = \beta_2 + \sum E(w_i e_i|\mathbf{x}) = \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 + \sum w_i 0 = \beta_2$$

Conditional on $\mathbf{x}$, which is equivalent to assuming $\mathbf{x}$ is given, the function $w_i = g(x_1, \ldots, x_N)$ is treated like a constant and is factored out in the third equality. Applying the law of iterated expectations, we find

$$E(b_2) = E_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] = E_{\mathbf{x}}(\beta_2) = \beta_2$$

The notation $E_{\mathbf{x}}(\ )$ means take the expected value treating $\mathbf{x}$ as random. In this case, that is not difficult because β_2 is a constant, nonrandom parameter. The least squares estimator is unbiased, both conditional on $\mathbf{x}$ and unconditionally, under strict exogeneity.

The derivation of the variance of the least squares estimator changes in a similar way. First find the variance of b_2 given $\mathbf{x}$.

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= E\left[\left(b_2 - E(b_2|\mathbf{x})\right)^2 \middle| \mathbf{x}\right] = E\left[(\beta_2 + \sum w_i e_i - \beta_2)^2 \middle| \mathbf{x}\right] = E\left[\left(\sum w_i e_i\right)^2 \middle| \mathbf{x}\right] \\
&= E\left[\left(\sum w_i^2 e_i^2 + \sum\sum_{i \neq j} w_i w_j e_i e_j\right) \middle| \mathbf{x}\right] = \sum w_i^2 E(e_i^2|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j E(e_i e_j|\mathbf{x}) \\
&= \sigma^2 \sum w_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

The variance of b_2 given $\mathbf{x}$ is exactly the same as when $\mathbf{x}$ was assumed random and statistically independent of the random errors. Now find the variance of b_2 using the variance decomposition from the Probability Primer equation (P.29). For two random variables X and Y,

$$\text{var}(Y) = \text{var}_X[E(Y|X)] + E_X[\text{var}(Y|X)]$$

Letting $Y = b_2$ and $X = \mathbf{x}$, we have

$$\text{var}(b_2) = \text{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] + E_{\mathbf{x}}\left[\text{var}(b_2|\mathbf{x})\right] = \text{var}_{\mathbf{x}}(\beta_2) + E_{\mathbf{x}}\left[\frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right] = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum (x_i - \bar{x})^2}\right]$$

since $\text{var}_{\mathbf{x}}(\beta_2) = 0$. This is exactly the same result as in the case in which x_j and e_i are statistically independent.

2G.4 Random Sampling

In the case of random sampling, data pairs (y_i, x_i) are *iid*, and the strict exogeneity assumption reduces to $E(e_i|x_i) = 0$. The results in the previous section hold in exactly the same way because it is still true that $E(e_i|\mathbf{x}) = 0$.

Appendix 2H Monte Carlo Simulation

The statistical properties of the least squares estimators are well known if the assumptions in Section 2.1 hold. In fact, we know that the least squares estimators are the best linear unbiased estimators of the regression parameters under these assumptions. And if the random errors are normal, then we know that, given **x**, the estimators themselves have normal distributions in **repeated experimental trials**. The meaning of "repeated trials" is difficult to grasp. **Monte Carlo** simulation experiments use random number generators to replicate the random way that data are obtained. In Monte Carlo simulations, we specify a **data generation process** and create samples of artificial data. Then, we "try out" estimation methods on the data we have created. We create **many** samples of size N and examine the **repeated sampling properties** of the estimators. In this way, we can study how statistical procedures behave under ideal, as well as not so ideal, conditions. This is important because economic, business, and social science data are not always (indeed, not usually) as nice as the assumptions we make.

The DGP for the simple linear regression model is given by

$$y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

Each value of the dependent variable y_i is obtained, or generated, by adding a random error e_i to the regression function $E(y_i|x_i)$. To simulate values of y_i, we create values for the systematic portion of the regression relationship $E(y_i|x_i)$ and add to it the random error e_i. This is analogous to a physical experiment in which variable factors are set at fixed levels and the experiment run. The outcome is different in each experimental trial because of random uncontrolled errors.

2H.1 The Regression Function

The regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ is the systematic portion of the regression relationship. To create these values we must select the following:

1. *A sample size N.* From the discussion in Section 2.4.4, we know that the larger the sample size is, the greater is the precision of estimation of the least squares estimators b_1 and b_2. Following the numerical examples in the book, we choose $N = 40$. This is not a large sample, but assuming SR1–SR5 are true, the least squares estimators' properties hold for any sample of size $N > 2$ in the simple regression model. In more complex situations, varying the sample size to see how estimators perform is an important ingredient of the simulation.
2. *We must choose x_i values.* For simplicity, we initially assume values of the explanatory variable that are fixed in repeated experimental trials. Following the depiction in Figure 2.1,[15] we set the values $x_1, x_2, \ldots, x_{20} = 10$ and $x_{21}, x_{22}, \ldots, x_{40} = 20$, using the chapter assumption that x is measured in hundreds. Does it matter how we choose the x_i values? Yes, it does. The variances and covariances of the least squares estimators depend on the variation in x_i, $\sum(x_i - \bar{x})^2$, how far the values are from 0, as measured by $\sum x_i^2$, and on the sample mean $\bar{x}$. Thus, if the values x_i change, the precision of estimation of the least squares estimators will change.

[15] This design is used in Briand, G. & Hill, R. C. (2013). Teaching Basic Econometric Concepts using Monte Carlo Simulations in Excel, *International Review of Economics Education*, 12(1), 60–79.

3. *We must choose* β_1 *and* β_2. Interestingly, for the least squares estimator under assumptions SR1–SR5, the actual magnitudes of these parameters do not matter a great deal. The estimator variances and covariances do not depend on them. The difference between the least squares estimator and the true parameter value, $E(b_2) - \beta_2$ given in (2.13), does not depend on the magnitude of β_2, only on the x_i values and the random errors e_i. To roughly parallel the regression results we obtained in Figure 2.10, we set $\beta_1 = 100$ and $\beta_2 = 10$.

Given the values above we can create $N = 40$ values $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. These values are

$$E(y_i|x_i = 10) = 100 + 10x_i = 100 + 10 \times 10 = 200, \qquad i = 1, \ldots, 20$$
$$E(y_i|x_i = 20) = 100 + 10x_i = 100 + 10 \times 20 = 300, \qquad i = 21, \ldots, 40$$

2H.2 The Random Error

To be consistent with assumptions SR2–SR4, the random errors should have mean zero, constant variance $\text{var}(e_i|x_i) = \sigma^2$ and be uncorrelated with one another, so that $\text{cov}(e_i, e_j|\mathbf{x}) = 0$. Researchers in the field of numerical analysis have studied how to simulate random numbers from a variety of probability distributions, such as the normal distribution. Of course, the computer-generated numbers cannot be truly random, because they are generated by a computer code. The random numbers created by computer software are "pseudorandom," in that they behave like random numbers. The numbers created will begin to recycle after about 2^{19937} values are drawn, using the so-called Mersenne Twister algorithm. Each software vendor uses its own version of a random number generator. Consequently, you should not expect to obtain exactly the same numbers that we have, and your replication will produce slightly different results, even though the major conclusions will be the same. See Appendix B.4 for a discussion of how random numbers are created.

Following assumption SR6, we assume the random error terms have a normal distribution with mean zero and a homoskedastic variance $\text{var}(e_i|x_i) = \sigma^2$. The variance σ^2 affects the precision of estimation through the variances and covariances of the least squares estimators in (2.14)–(2.16). The bigger the value of σ^2, the bigger the variances and covariances of the least squares estimators, and the more spread out the probability distribution of the estimators, as shown in Figure 2.11. We choose $\text{var}(e_i|x_i) = \sigma^2 = 2500$, which also means that $\text{var}(y_i|x_i) = \sigma^2 = 2500$.

2H.3 Theoretically True Values

Using the values above, we plot the theoretically true *pdfs* for y_i in Figure 2H.1. The solid curve on the left is $N(200, 2500 = 50^2)$. The first 20 simulated observations will follow this *pdf*. The dashed curve on the right is $N(300, 2500 = 50^2)$, which is the *pdf* for the second 20 observations.

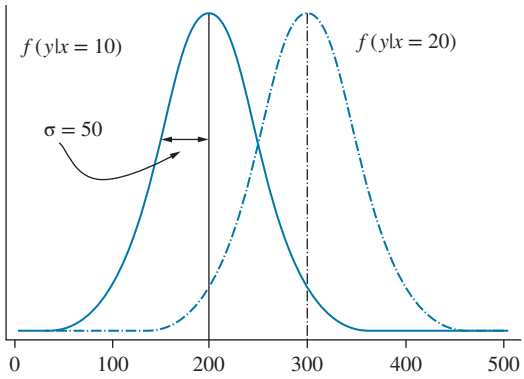


FIGURE 2H.1 **The true *pdfs* of the data.**

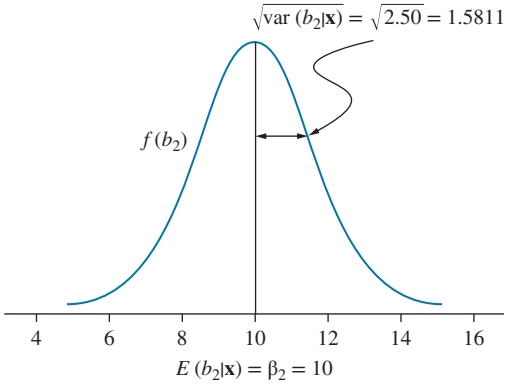


FIGURE 2H.2 **The true pdf of the estimator b_2.**

Given the parameter $\sigma^2 = 2500$ and the x_i values, we can compute the true conditional variances of the estimators:

$$\text{var}(b_1|\mathbf{x}) = \sigma^2\left[\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}\right] = 2500\left[\frac{10000}{40\times 1000}\right] = 625$$

$$\text{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} = \frac{2500}{1000} = 2.50$$

$$\text{cov}(b_1, b_2|\mathbf{x}) = \sigma^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] = 2500\left[\frac{-15}{1000}\right] = -37.50$$

The true standard deviation of b_2 is $\sqrt{\text{var}(b_2|\mathbf{x})} = \sqrt{2.50} = 1.5811$. The true *pdf* of $b_2|\mathbf{x}$ is $N(\beta_2 = 10, \text{var}(b_2|\mathbf{x}) = 2.5)$. Using the cumulative probabilities for the standard normal distribution in Statistical Table 1, we find that 98% of values from a normal distribution fall within 2.33 standard deviations of the mean. Applying this rule to the estimates b_2, we have

$$\beta_2 \pm 2.33 \times \sqrt{\text{var}(b_2|\mathbf{x})} = 10 \pm 2.33 \times 1.5811 = [6.316, 13.684]$$

We expect almost all values of b_2 (98% of them) to fall in the range 6.32–13.68. The plot of the true *pdf* of the estimator b_2 is shown in Figure 2H.2.

2H.4 Creating a Sample of Data

Most software will automatically create random values, z_i, from the standard normal distribution, $N(0, 1)$. To obtain a random value from a $N(0, \sigma^2)$ distribution, we multiply z_i by the standard deviation σ. That is, $e_i = \sigma \times z_i$. Given values z_i from the standard normal distribution, we obtain the $N = 40$ sample values from the chosen DGP as

$$y_i = E(y_i|x_i = 10) + e_i = 200 + 50 \times z_i \qquad i = 1, \ldots, 20$$
$$y_i = E(y_i|x_i = 20) + e_i = 300 + 50 \times z_i \qquad i = 21, \ldots, 40$$

One sample of data is in the data file *mc1_fixed_x*. Using these values, we obtain the least squares estimates. It is convenient to display the coefficient estimates and standard errors together, with the standard error reported below the coefficients:

$$\hat{y} = 127.2055 + 8.7325x$$
$$\text{(se)} \quad (23.3262) \quad (1.4753)$$

The estimate $\hat{\sigma} = 46.6525$. The estimated variances and covariance of b_1 and b_2 are $\widehat{\text{var}}(b_1) = 544.1133$, $\widehat{\text{var}}(b_2) = 2.1765$, and $\widehat{\text{cov}}(b_1, b_2) = -32.6468$.

For this one sample, the parameter estimates are reasonably near their true values. However, what happens in one sample does not prove anything. The repeated sampling properties of the least squares estimators are about what happens in many samples of data, from the same DGP.

2H.5 Monte Carlo Objectives

What do we hope to achieve with a Monte Carlo experiment? After the Monte Carlo experiment, we will have many least squares estimates. If we obtain $M = 10,000$ samples, we will have 10,000 estimates $b_{1,1}, \ldots, b_{1,M}$, 10,000 estimates $b_{2,1}, \ldots, b_{2,M}$, and 10,000 estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2$.

- We would like to verify that under SR1–SR5 the least squares estimators are unbiased. The estimator b_2 is unbiased if $E(b_2) = \beta_2$. Since an expected value is an average in many repeated experimental trials, we should observe that the average value of all the slope estimates, $\bar{b}_2 = \sum_{m=1}^{M} b_{2m}/M$, is close to $\beta_2 = 10$.
- We would like to verify that under SR1–SR5 the least squares estimators have sampling variances given by (2.14) and (2.16). The estimator variances measure the sampling variation in the estimates. The sampling variation of the estimates in the Monte Carlo simulation can be measured by their sample variance. For example, the sample variance of the estimates $b_{2,1}, \ldots, b_{2,M}$ is $s_{b_2}^2 = \sum_{m=1}^{M}\left(b_{2,m} - \bar{b}_2\right)^2/(M-1)$. This value should be close to $\text{var}(b_2) = 2.50$, and the standard deviation s_{b_2} should be close to the true standard deviation of the regression estimates 1.5811.
- We would like to verify that the estimator of the error variance (2.19) is an unbiased estimator of $\sigma^2 = 2500$, or that $\hat{\sigma}^2 = \sum_{m=1}^{M} \hat{\sigma}_m^2/M$ is close to the true value.
- Because we have assumed the random errors are normal, SR6, we expect the least squares estimates to have a normal distribution.

2H.6 Monte Carlo Results

The numerical results of the Monte Carlo experiment are shown in Table 2H.1. The averages (or "Sample Means") of the 10,000 Monte Carlo estimates are close to their true values.

For example, the average of the slope estimates is $\bar{b}_2 = \sum_{m=1}^{M} b_{2,m}/M = 10.0130$ compared to the true value $\beta_2 = 10$. The sample variance of the estimates $s_{b_2}^2 = \sum_{m=1}^{M}\left(b_{2,m} - \bar{b}_2\right)^2/(M-1) = 2.4691$ compared to the true value $\text{var}(b_2) = 2.50$. The standard deviation of the estimates is $s_{b_2} = 1.5713$ compared to the true standard deviation $\sqrt{\text{var}(b_2)} = \sqrt{2.50} = 1.5811$. The theoretical 1st and 99th percentiles of b_2 are [6.316, 13.684], which is reflected by the estimates [6.3268, 13.6576].

As for the normality of the estimates, we see from the histogram in Figure 2H.3 that the actual values follow the superimposed normal distribution very closely.[16]

TABLE 2H.1 **Summary of 10,000 Monte Carlo Samples**

	Mean	Variance	Std. Dev.	Minimum	Maximum	1st Pct.	99th Pct.
b_1 (100)	99.7463	613.4323	24.7676	12.1000	185.5361	42.2239	156.5996
b_2 (10)	10.0130	2.4691	1.5713	4.5881	16.5293	6.3268	13.6576
$\hat{\sigma}^2$ (2500)	2490.67	329964.7	574.4256	976.447	5078.383	1366.225	4035.681

[16] A normal distribution is symmetrical with no skewness, and for the estimates b_2 the skewness is −0.0027. A normal distribution has kurtosis of three, and for the estimates b_2 the kurtosis is 3.02. The Jarque–Bera test statistic that combines skewness and kurtosis measures is 0.1848 yielding a p-value of 0.91, meaning that we fail to reject the normality. See Appendix C.7.4 for a discussion of the Jarque–Bera test.

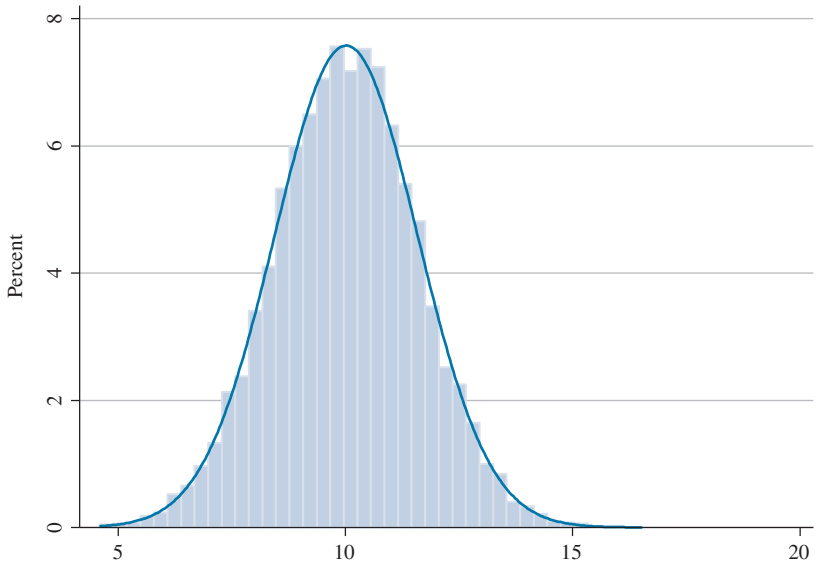


FIGURE 2H.3 **The sampling distribution of b_2 in 10,000 Monte Carlo samples when x is fixed in repeated trials.**

If you are replicating these results, some suggested exercises are as follows:

1. Test if mean of $\bar{b}_2$ is equal to β_2 using the test described in Appendix C.6.1.
2. Calculate the percentage of estimates falling in a given interval, such as between 8 and 9, and compare it to the probability based on the normal distribution.

2H.7 Random-x Monte Carlo Results

We used the "fixed-x" framework in the simulation results above. In each Monte Carlo sample, the x-values were $x_i = 10$ for the first 20 observations and $x_i = 20$ for the next 20 observations. Now we modify the experiment to the random-x case. The data generating equation remains $y_i = 100 + 10x_i + e_i$ with the random errors having a normal distribution with mean zero and standard deviation 50, $e_i \sim N(0, 50^2 = 2500)$. We randomly choose x-values from a normal distribution with mean $\mu_x = 15$ and standard deviation $\sigma_x = 1.6$, so $x \sim N(15, 1.6^2 = 2.56)$. We chose $\sigma_x = 1.6$ so that 99.73% of the **random-*x*** values fall between 10.2 and 19.8, which is similar in spirit to the fixed-x simulation in the previous section.

One sample of data is in the file *mc1_random_x*. Using these values, we obtain the least squares estimates and standard errors

$$\hat{y} = 116.7410 + 9.7628x$$
$$\text{(se)} \quad (84.7107) \quad (5.5248)$$

and the estimate $\hat{\sigma} = 51.3349$. The estimates are close to the true values.

The numerical results of the Monte Carlo experiment are shown in Table 2H.2. The averages (or "Sample Means") of the 10,000 Monte Carlo estimates are close to their true values.

For example, the average of the slope estimates is $\bar{b}_2 = \sum_{m=1}^{M} b_{2,m}/M = 10.0313$ compared to the true value $\beta_2 = 10$. In the random-x case, the true variance of the least squares estimator is

$$\text{var}(b_2) = \sigma^2 E\left[\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] = \frac{\sigma^2}{(N-3)\sigma_x^2} = \frac{2500}{(37)(2.56)} = 26.3936$$

Calculating the variance we use a special property resulting from the normality of x. When x is normally distributed $N(\mu_x, \sigma_x^2)$ the unbiased estimator of σ_x^2 is $s_x^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/(N-1)$.

TABLE 2H.2 **Summary of 10,000 Random-x Monte Carlo Samples**

	Mean	Var.	Std. Dev.	Min.	Max.	1st Pct.	99th Pct.
b_1 (100)	99.4344	6091.4412	78.0477	−196.8826	405.8328	−83.1178	283.8266
b_2 (10)	10.0313	26.8503	5.1817	−10.4358	29.3168	−2.2196	22.3479
$\widehat{\text{var}}(b_2)$ (26.3936)	26.5223	78.9348	8.8845	7.8710	91.1388	11.8325	54.0177
$\hat{\sigma}^2$ (2500)	2498.4332	332622.6	576.7344	809.474	5028.047	1366.957	4056.279

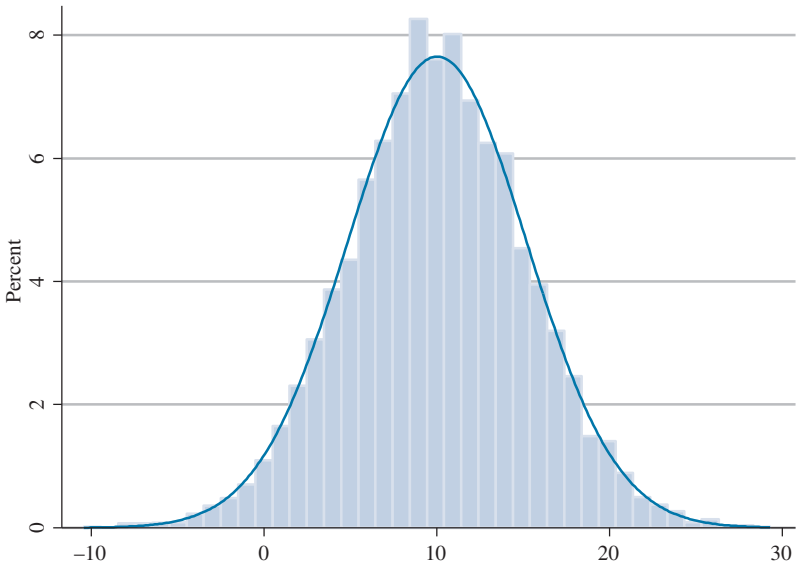


FIGURE 2H.4 **The sampling distribution of b_2 in 10,000 Monte Carlo samples when x is random in repeated trials.**

In Appendix C.7.1 we use the fact that $(N-1)s_x^2/\sigma_x^2 \sim \chi^2_{(N-1)}$. This implies that $V = \sum_{i=1}^{N}(x_i - \bar{x})^2 \sim \sigma_x^2 \chi^2_{(N-1)}$. Using the properties of the inverse chi-square distribution $E(1/V) = E\left[1/\sum_{i=1}^{N}(x_i - \bar{x})^2\right] = 1/\left[(N-3)\sigma_x^2\right]$.[17] Note that the Monte Carlo mean of the estimated $\text{var}(b_2)$ is 26.5223, confirming that $\widehat{\text{var}}(b_2) = 2500/[37(2.56)] = 26.3936$ is an unbiased estimator even in the random-x case.

Recall, however, that in the random-x case the distribution of the least squares estimator b_2 is not normal. The histogram of the 10,000 Monte Carlo estimates is shown in Figure 2H.4. It is symmetrical but there are too many central values, and the peak is too high. Statistically we can reject that this distribution is normal.[18]

If you are replicating these results, some suggested exercises are as follows:

1. Test if mean of $\bar{b}_2$ is equal to β_2 using the test described in Appendix C.6.1.
2. Calculate the percentage of estimates falling in a given interval, such as between 8 and 9, and compare it with the probability based on the normal distribution.

[17] See Appendix B.3.6 and Appendix C.7.1 for the theory behind this result.

[18] A normal distribution is symmetrical with no skewness, and for the estimates b_2 the skewness is −0.001. A normal distribution has kurtosis of three, and for the estimates b_2 the kurtosis is 3.14. The Jarque–Bera test statistic that combines skewness and kurtosis measures is 8.32 yielding a p-value of 0.016, meaning that we reject the hypothesis of normality at the 5% level of significance. See Appendix C.7.4 for a discussion of the Jarque–Bera test.

CHAPTER 3

Interval Estimation and Hypothesis Testing

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Discuss how ''sampling theory'' relates to interval estimation and hypothesis testing.
2. Explain why it is important for statistical inference that given **x** the least squares estimators b_1 and b_2 are normally distributed random variables.
3. Explain the ''level of confidence'' of an interval estimator, and exactly what it means in a sampling context, and give an example.
4. Explain the difference between an interval estimator and an interval estimate. Explain how to interpret an interval estimate.
5. Explain the terms null hypothesis, alternative hypothesis, and rejection region, giving an example and a sketch of the rejection region.
6. Explain the logic of a statistical test, including why it is important that a test statistic has a known probability distribution if the null hypothesis is true.
7. Explain the term *p*-value and how to use a *p*-value to determine the outcome of a hypothesis test; provide a sketch showing a *p*-value.
8. Explain the difference between one-tail and two-tail tests. Explain, intuitively, how to choose the rejection region for a one-tail test.
9. Explain Type I error and illustrate it in a sketch. Define the level of significance of a test.
10. Explain the difference between economic and statistical significance.
11. Explain how to choose what goes in the null hypothesis and what goes in the alternative hypothesis.

KEYWORDS

alternative hypothesis
confidence intervals
critical value
degrees of freedom
hypotheses
hypothesis testing
inference
interval estimation
level of significance
linear combination of parameters
linear hypothesis
null hypothesis
one-tail tests
pivotal statistic
point estimates
probability value
p-value
rejection region
test of significance
test statistic
two-tail tests
Type I error
Type II error

In Chapter 2, we used the least squares estimators to develop **point estimates** for the parameters in the simple linear regression model. These estimates represent an **inference** about the regression function $E(y|x) = \beta_1 + \beta_2 x$ describing a relationship between economic variables. *Infer* means "to conclude by reasoning from something known or assumed." This dictionary definition describes statistical inference as well. We have assumed a relationship between economic variables and made various assumptions (SR1–SR5) about the regression model. Based on these assumptions, and given empirical estimates of regression parameters, we want to make inferences about the population from which the data were obtained.

In this chapter, we introduce additional tools of statistical inference: **interval estimation** and **hypothesis testing**. Interval estimation is a procedure for creating ranges of values, sometimes called **confidence intervals**, in which the unknown parameters are likely to be located. Hypothesis tests are procedures for comparing conjectures that we might have about the regression parameters to the parameter estimates we have obtained from a sample of data. Hypothesis tests allow us to say that the data are compatible, or are not compatible, with a particular conjecture or hypothesis.

The procedures for hypothesis testing and interval estimation depend very heavily on assumption SR6 of the simple linear regression model and the resulting conditional normality of the least squares estimators. If assumption SR6 does not hold, then the sample size must be sufficiently large so that the distributions of the least squares estimators are *approximately* normal. In this case, the procedures we develop in this chapter can be used but are also approximate. In developing the procedures in this chapter, we will be using the "Student's" t-distribution. You may want to refresh your memory about this distribution by reviewing Appendix B.3.7. In addition, it is sometimes helpful to see the concepts we are about to discuss in a simpler setting. In Appendix C, we examine statistical inference, interval estimation, and hypothesis testing in the context of estimating the mean of a normal population. You may want to review this material now or read it along with this chapter as we proceed.

3.1 Interval Estimation

In Chapter 2, in Example 2.4, we estimated that household food expenditure would rise by \$10.21 given a \$100 increase in weekly income. The estimate $b_2 = 10.21$ is a *point* estimate of the unknown population parameter β_2 in the regression model. Interval estimation proposes a range of values in which the true parameter β_2 is likely to fall. Providing a range of values gives a sense of what the parameter value might be, and the precision with which we have estimated it. Such intervals are often called **confidence intervals**. We prefer to call them **interval estimates** because the term "confidence" is widely misunderstood and misused. As we will see, our confidence is in the procedure we use to obtain the intervals, not in the intervals themselves. This is consistent with how we assessed the properties of the least squares estimators in Chapter 2.

3.1.1 The t-Distribution

Let us assume that assumptions SR1–SR6 hold for the simple linear regression model. In this case, we know that given $\mathbf{x}$ the least squares estimators b_1 and b_2 have normal distributions, as discussed in Section 2.6. For example, the normal distribution of b_2, the least squares estimator of β_2, is

$$b_2|\mathbf{x} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

A standardized normal random variable is obtained from b_2 by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \sim N(0, 1) \tag{3.1}$$

The standardized random variable Z is normally distributed with mean 0 and variance 1. By standardizing the conditional normal distribution of $b_2|\mathbf{x}$, we find a statistic Z whose $N(0, 1)$ sampling distribution does not depend on any unknown parameters or on $\mathbf{x}$! Such statistics are called **pivotal**, and this means that when making probability statements about Z we do not have to worry about whether $\mathbf{x}$ is fixed or random. Using a table of normal probabilities (Statistical Table 1) we know that

$$P(-1.96 \le Z \le 1.96) = 0.95$$

Substituting (3.1) into this expression, we obtain

$$P\left(-1.96 \le \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \le 1.96\right) = 0.95$$

Rearranging gives us

$$P\left(b_2 - 1.96\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2} \le \beta_2 \le b_2 + 1.96\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}\right) = 0.95$$

This defines an interval that has probability 0.95 of containing the parameter β_2. The two endpoints $\left(b_2 \pm 1.96\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}\right)$ provide an interval estimator. If we construct intervals this way using all possible samples of size N from a population, then 95% of the intervals will contain the true parameter β_2. This easy derivation of an interval estimator is based on both assumption SR6 *and* our knowing the variance of the error term σ^2.

Although we do not know the value of σ^2, we can estimate it. The least squares residuals are $\hat{e}_i = y_i - b_1 - b_2 x_i$, and our estimator of σ^2 is $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-2)$. Replacing σ^2 by $\hat{\sigma}^2$ in (3.1) creates a random variable we can work with, but this substitution changes the probability distribution from standard normal to a t-distribution with $N-2$ **degrees of freedom**,

$$t = \frac{b_2 - \beta_2}{\sqrt{\hat{\sigma}^2 / \sum(x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\sqrt{\widehat{\text{var}}(b_2)}} = \frac{b_2 - \beta_2}{\text{se}(b_2)} \sim t_{(N-2)} \tag{3.2}$$

The ratio $t = (b_2 - \beta_2)/\text{se}(b_2)$ has a t-distribution with $N-2$ degrees of freedom, which we denote as $t \sim t_{(N-2)}$. By standardizing the conditional normal distribution of $b_2|\mathbf{x}$ and inserting the estimator $\hat{\sigma}^2$, we find a statistic t whose $t_{(N-2)}$ sampling distribution does not depend on any unknown parameters or on $\mathbf{x}$! It too is a **pivotal statistic**, and when making probability statements with a t-statistic, we do not have to worry about whether $\mathbf{x}$ is fixed or random. A similar result holds for b_1, so in general we can say, if assumptions SR1–SR6 hold in the simple linear regression model, then

$$t = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-2)} \quad \text{for} \quad k = 1, 2 \tag{3.3}$$

This equation will be the basis for interval estimation and hypothesis testing in the simple linear regression model. The statistical argument of how we go from (3.1) to (3.2) is in Appendix 3A.

When working with the t-distribution, remember that it is a bell-shaped curve centered at zero. It looks like the standard normal distribution, except that it is more spread out, with a larger variance and thicker tails. The shape of the t-distribution is controlled by a single parameter called the **degrees of freedom**, often abbreviated as df. We use the notation $t_{(m)}$ to specify a t-distribution with m degrees of freedom. In Statistical Table 2, there are percentile values of the t-distribution for various degrees of freedom. For m degrees of freedom, the 95th percentile of the t-distribution is denoted $t_{(0.95,\, m)}$. This value has the property that 0.95 of the probability falls to its left, so $P\left[t_{(m)} \le t_{(0.95,\, m)}\right] = 0.95$. For example, if the degrees of freedom are $m = 20$, then, from Statistical Table 2, $t_{(0.95,\, 20)} = 1.725$. Should you encounter a problem requiring percentiles

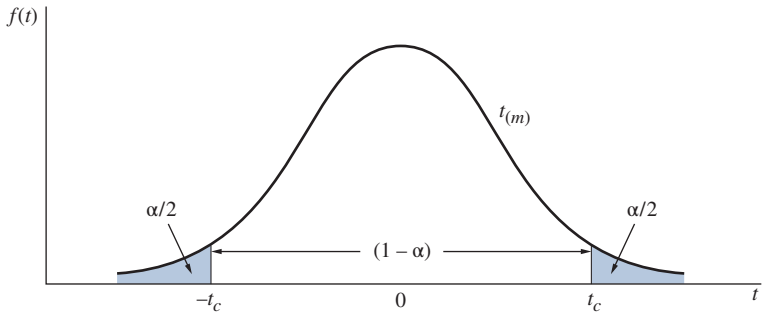


FIGURE 3.1 **Critical values from a *t*-distribution.**

that we do not give, you can interpolate for an approximate answer or use your computer software to obtain an exact value.

3.1.2 Obtaining Interval Estimates

From Statistical Table 2, we can find a "**critical value**" t_c from a t-distribution such that $P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$, where α is a probability often taken to be $\alpha = 0.01$ or $\alpha = 0.05$. The critical value t_c for degrees of freedom m is the percentile value $t_{(1-\alpha/2,\, m)}$. The values t_c and $-t_c$ are depicted in Figure 3.1.

Each shaded "tail" area contains $\alpha/2$ of the probability, so that $1 - \alpha$ of the probability is contained in the center portion. Consequently, we can make the probability statement

$$P(-t_c \leq t \leq t_c) = 1 - \alpha \tag{3.4}$$

For a 95% confidence interval, the critical values define a central region of the t-distribution containing probability $1 - \alpha = 0.95$. This leaves probability $\alpha = 0.05$ divided equally between the two tails, so that $\alpha/2 = 0.025$. Then the critical value $t_c = t_{(1-0.025,\, m)} = t_{(0.975,\, m)}$. In the simple regression model, the degrees of freedom are $m = N - 2$, so expression (3.4) becomes

$$P\left[-t_{(0.975,\, N-2)} \leq t \leq t_{(0.975,\, N-2)}\right] = 0.95$$

We find the percentile values $t_{(0.975,\, N-2)}$ in Statistical Table 2.

Now, let us see how we can put all these bits together to create a procedure for interval estimation. Substitute t from (3.3) into (3.4) to obtain

$$P\left[-t_c \leq \frac{b_k - \beta_k}{\text{se}(b_k)} \leq t_c\right] = 1 - \alpha$$

Rearrange this expression to obtain

$$P\left[b_k - t_c \text{se}(b_k) \leq \beta_k \leq b_k + t_c \text{se}(b_k)\right] = 1 - \alpha \tag{3.5}$$

The interval endpoints $b_k - t_c \text{se}(b_k)$ and $b_k + t_c \text{se}(b_k)$ are random because they vary from sample to sample. These endpoints define an **interval estimator** of β_k. The probability statement in (3.5) says that the interval $b_k \pm t_c \text{se}(b_k)$ has probability $1 - \alpha$ of containing the true but unknown parameter β_k.

When b_k and $\text{se}(b_k)$ in (3.5) are estimated values (numbers), based on a given sample of data, then $b_k \pm t_c \text{se}(b_k)$ is called a $100(1 - \alpha)\%$ **interval estimate** of β_k. Equivalently, it is called a $100(1 - \alpha)\%$ **confidence interval**. Usually, $\alpha = 0.01$ or $\alpha = 0.05$, so that we obtain a 99% confidence interval or a 95% confidence interval.

The interpretation of confidence intervals requires a great deal of care. The properties of the interval estimation procedure are based on the notion of sampling. If we collect all possible

samples of size N from a population, compute the least squares estimate b_k and its standard error $\text{se}(b_k)$ for each sample, and then construct the interval estimate $b_k \pm t_c \text{se}(b_k)$ for each sample, then $100(1-\alpha)\%$ of all the intervals constructed would contain the true parameter β_k. In Appendix 3C, we carry out a Monte Carlo simulation to demonstrate this sampling property.

Any *one* interval estimate, based on one sample of data, may or may not contain the true parameter β_k, and because β_k is unknown, we will never know whether it does or does not. When "confidence intervals" are discussed, remember that our confidence is in the *procedure* used to construct the interval estimate; it is *not* in any one interval estimate calculated from a sample of data.

EXAMPLE 3.1 | Interval Estimate for Food Expenditure Data

For the food expenditure data, $N = 40$ and the degrees of freedom are $N - 2 = 38$. For a 95% confidence interval, $\alpha = 0.05$. The critical value $t_c = t_{(1-\alpha/2,\, N-2)} = t_{(0.975,\, 38)} = 2.024$ is the 97.5 percentile from the t-distribution with 38 degrees of freedom. For β_2, the probability statement in (3.5) becomes

$$P\left[b_2 - 2.024\text{se}(b_2) \le \beta_2 \le b_2 + 2.024\text{se}(b_2)\right] = 0.95 \tag{3.6}$$

To construct an interval estimate for β_2, we use the least squares estimate $b_2 = 10.21$ and its standard error

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2)} = \sqrt{4.38} = 2.09$$

Substituting these values into (3.6), we obtain a "95% confidence interval estimate" for β_2:

$$b_2 \pm t_c \text{se}(b_2) = 10.21 \pm 2.024(2.09) = [5.97, 14.45]$$

That is, we estimate "with 95% confidence" that from an additional \$100 of weekly income households will spend between \$5.97 and \$14.45 on food.

Is β_2 actually in the interval [5.97, 14.45]? We do not know, and we will never know. What we *do* know is that when the procedure we used is applied to all possible samples of data from the same population, then 95% of all the interval estimates constructed using this procedure will contain the true parameter. The interval estimation procedure "works" 95% of the time. What we can say about the interval estimate based on our one sample is that, given the reliability of the procedure, we would be "surprised" if β_2 is not in the interval [5.97, 14.45].

What is the usefulness of an interval estimate of β_2? When reporting regression results, we always give a point estimate, such as $b_2 = 10.21$. However, the point estimate alone gives no sense of its reliability. Thus, we might also report an interval estimate. Interval estimates incorporate both the point estimate and the standard error of the estimate, which is a measure of the variability of the least squares estimator. The interval estimate includes an allowance for the sample size as well because for lower degrees of freedom the t-distribution critical value t_c is larger. If an interval estimate is wide (implying a large standard error), it suggests that there is not much information in the sample about β_2. If an interval estimate is narrow, it suggests that we have learned more about β_2.

What is "wide" and what is "narrow" depend on the problem at hand. For example, in our model, $b_2 = 10.21$ is an estimate of how much weekly household food expenditure will rise given a \$100 increase in weekly household income. A CEO of a supermarket chain can use this estimate to plan future store capacity requirements, given forecasts of income growth in an area. However, no decision will be based on this one number alone. The prudent CEO will carry out a sensitivity analysis by considering values of β_2 around 10.21. The question is "Which values?" One answer is provided by the interval estimate [5.97, 14.45]. Though β_2 may or may not be in this interval, the CEO knows that the procedure used to obtain the interval estimate "works" 95% of the time. If varying β_2 within the interval has drastic consequences on company sales and profits, then the CEO may conclude that there is insufficient evidence upon which to make a decision and order a new and larger data sample.

3.1.3 The Sampling Context

In Section 2.4.3, we illustrated the sampling properties of the least squares estimators using 10 data samples. Each sample of size $N = 40$ includes households with the same incomes as in Table 2.1 but with food expenditures that vary. These hypothetical data are in the data file *table2_2*. In Table 3.1, we present the OLS estimates, the estimates of σ^2, and the coefficient standard errors from each sample. Note the sampling variation illustrated by these estimates. The

TABLE 3.1 Least Squares Estimates from 10 Hypothetical Random Samples

Sample	b_1	$\text{se}(b_1)$	b_2	$\text{se}(b_2)$	$\hat{\sigma}^2$
1	93.64	31.73	8.24	1.53	4282.13
2	91.62	31.37	8.90	1.51	4184.79
3	126.76	48.08	6.59	2.32	9828.47
4	55.98	45.89	11.23	2.21	8953.17
5	87.26	42.57	9.14	2.05	7705.72
6	122.55	42.65	6.80	2.06	7735.38
7	91.95	42.14	9.84	2.03	7549.82
8	72.48	34.04	10.50	1.64	4928.44
9	90.34	36.69	8.75	1.77	5724.08
10	128.55	50.14	6.99	2.42	10691.61

TABLE 3.2 Interval Estimates from 10 Hypothetical Random Samples

Sample	$b_1 - t_c\text{se}(b_1)$	$b_1 + t_c\text{se}(b_1)$	$b_2 - t_c\text{se}(b_2)$	$b_2 + t_c\text{se}(b_2)$
1	29.40	157.89	5.14	11.34
2	28.12	155.13	5.84	11.96
3	29.44	224.09	1.90	11.29
4	−36.91	148.87	6.75	15.71
5	1.08	173.43	4.98	13.29
6	36.21	208.89	2.63	10.96
7	6.65	177.25	5.73	13.95
8	3.56	141.40	7.18	13.82
9	16.07	164.62	5.17	12.33
10	27.04	230.06	2.09	11.88

variation is due to the fact that in each sample household food expenditures are different. The 95% confidence intervals for the parameters β_1 and β_2 are given in Table 3.2 for the same samples.

Sampling variability causes the center of each of the interval estimates to change with the values of the least squares estimates, and it causes the widths of the intervals to change with the standard errors. If we ask the question "How many of these intervals contain the true parameters, and which ones are they?" we must answer that we do not know. But since 95% of all interval estimates constructed this way contain the true parameter values, we would expect perhaps 9 or 10 of these intervals to contain the true but unknown parameters.

Note the difference between point estimation and interval estimation. We have used the least squares estimators to obtain point estimates of unknown parameters. The estimated variance $\widehat{\text{var}}(b_k)$, for $k = 1$ or 2, and its square root $\sqrt{\widehat{\text{var}}(b_k)} = \text{se}(b_k)$ provide information about the sampling variability of the least squares estimator from one sample to another. Interval estimators are a convenient way to report regression results because they combine point estimation with a measure of sampling variability to provide a range of values in which the unknown parameters might fall. When the sampling variability of the least squares estimator is relatively small, then the interval estimates will be relatively narrow, implying that the least squares estimates are "reliable." If the least squares estimators suffer from large sampling variability, then the interval estimates will be wide, implying that the least squares estimates are "unreliable."

3.2 Hypothesis Tests

Many business and economic decision problems require a judgment as to whether or not a parameter is a specific value. In the food expenditure example, it may make a good deal of difference for decision purposes whether β_2 is greater than 10, indicating that a \$100 increase in income will increase expenditure on food by more than \$10. In addition, based on economic theory, we believe that β_2 should be positive. One check of our data and model is whether this theoretical proposition is supported by the data.

Hypothesis testing procedures compare a conjecture we have about a population to the information contained in a sample of data. Given an economic and statistical model, **hypotheses** are formed about economic behavior. These hypotheses are then represented as statements about model parameters. Hypothesis tests use the information about a parameter that is contained in a sample of data, its least squares point estimate, and its standard error to draw a conclusion about the hypothesis.

In each and every hypothesis test, five ingredients must be present:

Components of Hypothesis Tests

1. A null hypothesis H_0
2. An alternative hypothesis H_1
3. A test statistic
4. A rejection region
5. A conclusion

3.2.1 The Null Hypothesis

The **null hypothesis**, which is denoted by H_0 (*H-naught*), specifies a value for a regression parameter, which for generality we denote as β_k, for $k = 1$ or 2. The null hypothesis is stated as $H_0 : \beta_k = c$, where c is a constant, and is an important value in the context of a specific regression model. A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true, in which case we *reject* the null hypothesis.

3.2.2 The Alternative Hypothesis

Paired with every null hypothesis is a logical **alternative hypothesis** H_1 that we will accept if the null hypothesis is rejected. The alternative hypothesis is flexible and depends, to some extent, on economic theory. For the null hypothesis $H_0 : \beta_k = c$, the three possible alternative hypotheses are as follows:

- $H_1 : \beta_k > c$. Rejecting the null hypothesis that $\beta_k = c$ leads us to accept the conclusion that $\beta_k > c$. Inequality alternative hypotheses are widely used in economics because economic theory frequently provides information about the *signs* of relationships between variables. For example, in the food expenditure example, we might well test the null hypothesis $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 > 0$ because economic theory strongly suggests that necessities such as food are normal goods and that food expenditure will rise if income increases.
- $H_1 : \beta_k < c$. Rejecting the null hypothesis that $\beta_k = c$ in this case leads us to accept the conclusion that $\beta_k < c$.
- $H_1 : \beta_k \neq c$. Rejecting the null hypothesis that $\beta_k = c$ in this case leads us to accept the conclusion that β_k takes a value either larger or smaller than c.

3.2.3 The Test Statistic

The sample information about the null hypothesis is embodied in the sample value of a **test statistic**. Based on the value of a test statistic, we decide either to reject the null hypothesis or not to reject it. A test statistic has a special characteristic: its probability distribution is completely *known* when the null hypothesis is true, and it has some *other* distribution if the null hypothesis is not true.

It all starts with the key result in (3.3), $t = (b_k - \beta_k)/\text{se}(b_k) \sim t_{(N-2)}$. **If** the null hypothesis $H_0 : \beta_k = c$ is *true*, **then** we can substitute c for β_k and it follows that

$$t = \frac{b_k - c}{\text{se}(b_k)} \sim t_{(N-2)} \tag{3.7}$$

If the null hypothesis is *not true*, then the t-statistic in (3.7) does *not* have a t-distribution with $N - 2$ degrees of freedom. This point is elaborated in Appendix 3B.

3.2.4 The Rejection Region

The **rejection region** depends on the form of the alternative. It is the range of values of the test statistic that leads to *rejection* of the null hypothesis. It is possible to construct a rejection region only if we have

- A test statistic whose distribution is known when the null hypothesis is true
- An alternative hypothesis
- A level of significance

The rejection region consists of values that are *unlikely* and that have low probability of occurring when the null hypothesis is true. The chain of logic is "If a value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus, it is unlikely that the null hypothesis is true." If the alternative hypothesis is true, then values of the test statistic will tend to be unusually large or unusually small. The terms "large" and "small" are determined by choosing a probability α, called the **level of significance** of the test, which provides a meaning for "an *unlikely* event." The level of significance of the test α is usually chosen to be 0.01, 0.05, or 0.10.

Remark

When no other specific choice is made, economists and statisticians often use a significance level of 0.05. That is, an occurrence "one time in twenty" is regarded as an unusual or improbable event by chance. This threshold for statistical significance is clung to as the Holy Grail but in reality is simply a historical precedent based on quotes by Sir Ronald Fisher who promoted the standard that t-values larger than two be regarded as significant.[1] A stronger threshold for significance, such as "one time in a hundred," or 0.01, might make more sense. The importance of the topic is quickly evident with a web search. The issues are discussed in *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, by Stephen T. Ziliak and Deirdre N. McCloskey, 2008, The University of Michigan Press.

If we reject the null hypothesis when it is true, then we commit what is called a **Type I error**. The level of significance of a test *is* the probability of committing a Type I error, so

[1] Mark Kelly (2013) "Emily Dickinson and monkeys on the stair. Or: What is the significance of the 5% significance level," *Significance*, Vol. 10(5), October, 21–22.

$P(\text{Type I error}) = \alpha$. Any time we reject a null hypothesis, it is possible that we have made such an error—there is no avoiding it. The good news is that we can specify the amount of Type I error we will tolerate by setting the level of significance α. If such an error is costly, then we make α small. If we do not reject a null hypothesis that is false, then we have committed a **Type II error**. In a real-world situation, we cannot control or calculate the probability of this type of error because it depends on the unknown true parameter β_k. For more about Type I and Type II errors, see Appendix C.6.9.

3.2.5 A Conclusion

When you have completed testing a hypothesis, you should state your conclusion. Do you reject the null hypothesis, or do you not reject the null hypothesis? As we will argue below, you should avoid saying that you "accept" the null hypothesis, which can be very misleading. Moreover, we urge you to make it standard practice to say what the conclusion means in the economic context of the problem you are working on and the economic significance of the finding. Statistical procedures are not ends in themselves. They are carried out for a reason and have meaning, which you should be able to explain.

3.3 Rejection Regions for Specific Alternatives

In this section, we hope to be very clear about the nature of the rejection rules for each of the three possible alternatives to the null hypothesis $H_0: \beta_k = c$. As noted in the previous section, to have a rejection region for a null hypothesis, we need a test statistic, which we have; it is given in (3.7). Second, we need a specific alternative, $\beta_k > c$, $\beta_k < c$, or $\beta_k \neq c$. Third, we need to specify the level of significance of the test. The level of significance of a test, α, is the probability that we reject the null hypothesis when it is actually true, which is called a Type I error.

3.3.1 One-Tail Tests with Alternative "Greater Than" (>)

When testing the null hypothesis $H_0: \beta_k = c$, if the *alternative* hypothesis $H_1: \beta_k > c$ is true, then the value of the t-statistic (3.7) tends to become larger than usual for the t-distribution. We will reject the null hypothesis if the test statistic is larger than the critical value for the level of significance α. The critical value that leaves probability α in the right tail is the $(1 - \alpha)$-percentile $t_{(1-\alpha, N-2)}$, as shown in Figure 3.2. For example, if $\alpha = 0.05$ and $N - 2 = 30$, then from Statistical Table 2, the critical value is the 95th percentile value $t_{(0.95, 30)} = 1.697$.

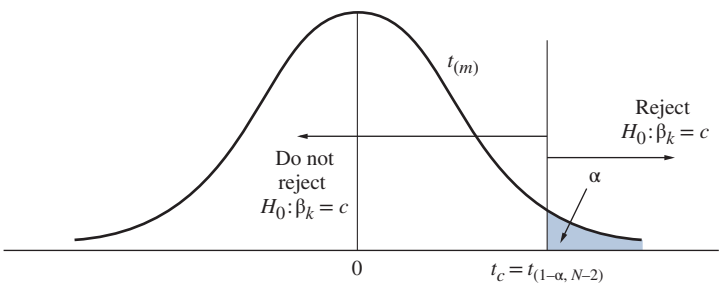


FIGURE 3.2 **Rejection region for a one-tail test of $H_0: \beta_k = c$ against $H_1: \beta_k > c$.**

The rejection rule is

> When testing the null hypothesis $H_0: \beta_k = c$ against the alternative hypothesis $H_1: \beta_k > c$, reject the null hypothesis and accept the alternative hypothesis if $t \geq t_{(1-\alpha, N-2)}$.

The test is called a "one-tail" test because unlikely values of the t-statistic fall only in one tail of the probability distribution. If the null hypothesis is true, then the test statistic (3.7) has a t-distribution, and its value would tend to fall in the center of the distribution, to the left of the critical value, where most of the probability is contained. The level of significance α is chosen so that if the null hypothesis is true, then the probability that the t-statistic value falls in the extreme right tail of the distribution is small; an event that is improbable and unlikely to occur by chance. If we obtain a test statistic value in the rejection region, we take it as evidence *against* the null hypothesis, leading us to conclude that the null hypothesis is unlikely to be true. Evidence against the null hypothesis is evidence in support of the alternative hypothesis. Thus, if we reject the null hypothesis then we conclude that the alternative is true.

If the null hypothesis $H_0: \beta_k = c$ is *true,* then the test statistic (3.7) has a t-distribution and its values fall in the nonrejection region with probability $1 - \alpha$. If $t < t_{(1-\alpha, N-2)}$, then there is no statistically significant evidence against the null hypothesis, and we do not reject it.

3.3.2 One-Tail Tests with Alternative "Less Than" (<)

If the alternative hypothesis $H_1: \beta_k < c$ is true, then the value of the t-statistic (3.7) tends to become smaller than usual for the t-distribution. We reject the null hypothesis if the test statistic is smaller than the critical value for the level of significance α. The critical value that leaves probability α in the left tail is the α-percentile $t_{(\alpha, N-2)}$, as shown in Figure 3.3.

When using Statistical Table 2 to locate critical values, recall that the t-distribution is symmetric about zero, so that the α-percentile $t_{(\alpha, N-2)}$ is the negative of the $(1 - \alpha)$-percentile $t_{(1-\alpha, N-2)}$. For example, if $\alpha = 0.05$ and $N - 2 = 20$, then from Statistical Table 2, the 95th percentile of the t-distribution is $t_{(0.95, 20)} = 1.725$ and the 5th percentile value is $t_{(0.05, 20)} = -1.725$.

The rejection rule is:

> When testing the null hypothesis $H_0: \beta_k = c$ against the alternative hypothesis $H_1: \beta_k < c$, reject the null hypothesis and accept the alternative hypothesis if $t \leq t_{(\alpha, N-2)}$.

The nonrejection region consists of t-statistic values greater than $t_{(\alpha, N-2)}$. When the null hypothesis is true, the probability of obtaining such a t-value is $1 - \alpha$, which is chosen to be large. Thus if $t > t_{(\alpha, N-2)}$ then do not reject $H_0: \beta_k = c$.

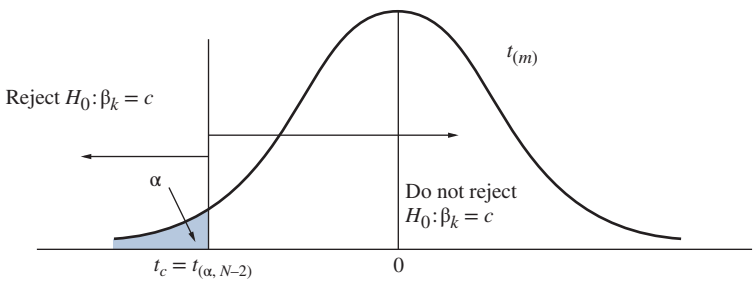


FIGURE 3.3 **The rejection region for a one-tail test of $H_0: \beta_k = c$ against $H_1: \beta_k < c$.**

Remembering where the rejection region is located may be facilitated by the following trick:

Memory Trick

The rejection region for a one-tail test is in the direction of the arrow in the alternative. If the alternative is >, then reject in the right tail. If the alternative is <, reject in the left tail.

3.3.3 Two-Tail Tests with Alternative "Not Equal To" (≠)

When testing the null hypothesis $H_0:\beta_k = c$, if the alternative hypothesis $H_1:\beta_k \neq c$ is true, then the value of the t-statistic (3.7) tends to become either larger *or* smaller than usual for the t-distribution. To have a test with the level of significance α, we define the critical values so that the probability of the t-statistic falling in either tail is $\alpha/2$. The left-tail critical value is the percentile $t_{(\alpha/2,\,N-2)}$ and the right-tail critical value is the percentile $t_{(1-\alpha/2,\,N-2)}$. We reject the null hypothesis that $H_0:\beta_k = c$ in favor of the alternative that $H_1:\beta_k \neq c$ if the test statistic $t \leq t_{(\alpha/2,\,N-2)}$ or $t \geq t_{(1-\alpha/2,\,N-2)}$, as shown in Figure 3.4. For example, if $\alpha = 0.05$ and $N - 2 = 30$, then $\alpha/2 = 0.025$ and the left-tail critical value is the 2.5-percentile value $t_{(0.025,\,30)} = -2.042$; the right-tail critical value is the 97.5-percentile $t_{(0.975,\,30)} = 2.042$. The right-tail critical value is found in Statistical Table 2, and the left-tail critical value is found using the symmetry of the t-distribution.

Since the rejection region is composed of portions of the t-distribution in the left and right tails, this test is called a **two-tail test**. When the null hypothesis is true, the probability of obtaining a value of the test statistic that falls in *either* tail area is "small." The sum of the tail probabilities is α. Sample values of the test statistic that are in the tail areas are incompatible with the null hypothesis and are evidence against the null hypothesis being true. On the other hand, if the null hypothesis $H_0:\beta_k = c$ is true, then the probability of obtaining a value of the test statistic t in the central nonrejection region is high. Sample values of the test statistic in the central nonrejection area are compatible with the null hypothesis and are not taken as evidence against the null hypothesis being true. Thus, the rejection rule is

When testing the null hypothesis $H_0:\beta_k = c$ against the alternative hypothesis $H_1:\beta_k \neq c$, reject the null hypothesis and accept the alternative hypothesis if $t \leq t_{(\alpha/2,\,N-2)}$ **or** if $t \geq t_{(1-\alpha/2,\,N-2)}$.

We do not reject the null hypothesis if $t_{(\alpha/2,\,N-2)} < t < t_{(1-\alpha/2,\,N-2)}$.

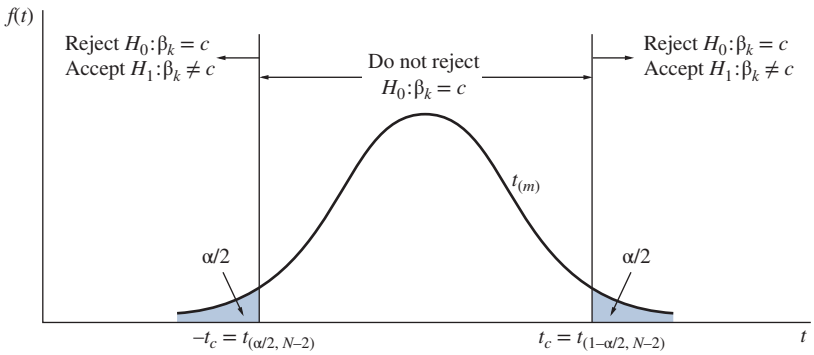


FIGURE 3.4 **Rejection region for a test of $H_0:\beta_k = c$ against $H_1:\beta_k \neq c$.**

3.4 Examples of Hypothesis Tests

We illustrate the mechanics of hypothesis testing using the food expenditure model. We give examples of right-tail, left-tail, and two-tail tests. In each case, we will follow a prescribed set of steps, closely following the list of required components for all hypothesis tests listed at the beginning of Section 3.2. A standard procedure for all hypothesis-testing problems and situations is

Step-by-Step Procedure for Testing Hypotheses

1. Determine the null and alternative hypotheses.
2. Specify the test statistic and its distribution if the null hypothesis is true.
3. Select α and determine the rejection region.
4. Calculate the sample value of the test statistic.
5. State your conclusion.

EXAMPLE 3.2 | Right-Tail Test of Significance

Usually, our first concern is whether there is a relationship between the variables, as we have specified in our model. If $\beta_2 = 0$, then there is no linear relationship between food expenditure and income. Economic theory suggests that food is a normal good and that as income increases food expenditure will also increase and thus that $\beta_2 > 0$. The least squares estimate of β_2 is $b_2 = 10.21$, which is certainly greater than zero. However, simply observing that the estimate has the correct sign does not constitute scientific proof. We want to determine whether there is convincing, or *significant*, statistical evidence that would lead us to conclude that $\beta_2 > 0$. When testing the null hypothesis that a parameter is zero, we are asking if the estimate b_2 is significantly different from zero, and the test is called a **test of significance**.

A statistical test procedure cannot prove the truth of a null hypothesis. When we fail to reject a null hypothesis, all the hypothesis test can establish is that the information in a sample of data is *compatible* with the null hypothesis. Conversely, a statistical test can lead us to *reject* the null hypothesis, with only a small probability α of rejecting the null hypothesis when it is actually true. Thus, rejecting a null hypothesis is a stronger conclusion than failing to reject it. For this reason, the null hypothesis is usually stated in such a way that if our theory is correct, then we will reject the null hypothesis. In our example, economic theory implies that there should be a positive relationship between income and food expenditure. We would like to establish that there is statistical evidence to support this theory using a hypothesis test. With this goal, we set up the null hypothesis that there is *no* relation between the variables, $H_0: \beta_2 = 0$. In the alternative hypothesis, we put the conjecture that we would like to establish, $H_1: \beta_2 > 0$. If we then reject the null hypothesis, we can make a direct statement, concluding that β_2 is positive, with only a small (α) probability that we are in error.

The steps of this hypothesis test are as follows:

1. The null hypothesis is $H_0: \beta_2 = 0$. The alternative hypothesis is $H_1: \beta_2 > 0$.
2. The test statistic is (3.7). In this case, $c = 0$, so $t = b_2/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true.
3. Let us select $\alpha = 0.05$. The critical value for the right-tail rejection region is the 95th percentile of the t-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.95,\,38)} = 1.686$. Thus, we will reject the null hypothesis if the calculated value of $t \geq 1.686$. If $t < 1.686$, we will not reject the null hypothesis.
4. Using the food expenditure data, we found that $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$. The value of the test statistic is
$$t = \frac{b_2}{\text{se}(b_2)} = \frac{10.21}{2.09} = 4.88$$
5. Since $t = 4.88 > 1.686$, we reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 > 0$. That is, we reject the hypothesis that there is no relationship between income and food expenditure and conclude that there is a *statistically significant* positive relationship between household income and food expenditure.

The last part of the conclusion is important. When you report your results to an audience, you will want to describe the

outcome of the test in the context of the problem you are investigating, not just in terms of Greek letters and symbols.

What if we had not been able to reject the null hypothesis in this example? Would we have concluded that economic theory is wrong and that there is no relationship between income and food expenditure? No. Remember that failing to reject a null hypothesis **does not** mean that the null hypothesis is true.

EXAMPLE 3.3 | Right-Tail Test of an Economic Hypothesis

Suppose that the economic profitability of a new supermarket depends on households spending more than \$5.50 out of each additional \$100 weekly income on food and that construction will not proceed unless there is strong evidence to this effect. In this case, the conjecture we want to establish, the one that will go in the alternative hypothesis, is that $\beta_2 > 5.5$. If $\beta_2 \leq 5.5$, then the supermarket will be unprofitable and the owners would not want to build it. The least squares estimate of β_2 is $b_2 = 10.21$, which is greater than 5.5. What we want to determine is whether there is convincing statistical evidence that would lead us to conclude, based on the available data, that $\beta_2 > 5.5$. This judgment is based on not only the estimate b_2 but also its precision as measured by $\text{se}(b_2)$.

What will the null hypothesis be? We have been stating null hypotheses as equalities, such as $\beta_2 = 5.5$. This null hypothesis is too limited because it is theoretically possible that $\beta_2 < 5.5$. It turns out that the hypothesis testing procedure for testing the null hypothesis that $H_0: \beta_2 \leq 5.5$ against the alternative hypothesis $H_1: \beta_2 > 5.5$ is *exactly the same* as testing $H_0: \beta_2 = 5.5$ against the alternative hypothesis $H_1: \beta_2 > 5.5$. The test statistic and rejection region are exactly the same. For a right-tail test, you can form the null hypothesis in either of these ways depending on the problem at hand.

The steps of this hypothesis test are as follows:

1. The null hypothesis is $H_0: \beta_2 \leq 5.5$. The alternative hypothesis is $H_1: \beta_2 > 5.5$.
2. The test statistic $t = (b_2 - 5.5)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true.
3. Let us select $\alpha = 0.01$. The critical value for the right-tail rejection region is the 99th percentile of the t-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.99,\,38)} = 2.429$. We will reject the null hypothesis if the calculated value of $t \geq 2.429$. If $t < 2.429$, we will not reject the null hypothesis.
4. Using the food expenditure data, $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$. The value of the test statistic is
$$t = \frac{b_2 - 5.5}{\text{se}(b_2)} = \frac{10.21 - 5.5}{2.09} = 2.25$$
5. Since $t = 2.25 < 2.429$, we do not reject the null hypothesis that $\beta_2 \leq 5.5$. We are *not* able to conclude that the new supermarket will be profitable and will not begin construction.

In this example, we have posed a situation where the choice of the level of significance α becomes of great importance. A construction project worth millions of dollars depends on having *convincing* evidence that households will spend more than \$5.50 out of each additional \$100 income on food. Although the "usual" choice is $\alpha = 0.05$, we have chosen a conservative value of $\alpha = 0.01$ because we seek a test that has a low chance of rejecting the null hypothesis when it is actually true. Recall that the level of significance of a test defines what we mean by an unlikely value of the test statistic. In this example, if the null hypothesis is true, then building the supermarket will be unprofitable. We want the probability of building an unprofitable market to be very small, and therefore, we want the probability of rejecting the null hypothesis when it is true to be very small. In each real-world situation, the choice of α must be made on an assessment of *risk* and the *consequences* of making an incorrect decision.

A CEO unwilling to make a decision based on the available evidence may well order a new and larger sample of data to be analyzed. Recall that as the sample size increases, the least squares estimator becomes more precise (as measured by estimator variance), and consequently, hypothesis tests become more powerful tools for statistical inference.

EXAMPLE 3.4 | Left-Tail Test of an Economic Hypothesis

For completeness, we will illustrate a test with the rejection region in the left tail. Consider the null hypothesis that $\beta_2 \geq 15$ and the alternative hypothesis $\beta_2 < 15$. Recall our memory trick for determining the location of the rejection region for a t-test. The rejection region is in the direction of the arrow $<$ in the alternative hypothesis. This fact tells us that the rejection region is in the left tail of the t-distribution. The steps of this hypothesis test are as follows:

1. The null hypothesis is $H_0: \beta_2 \geq 15$. The alternative hypothesis is $H_1: \beta_2 < 15$.
2. The test statistic $t = (b_2 - 15)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true.
3. Let us select $\alpha = 0.05$. The critical value for the left-tail rejection region is the 5th percentile of the t-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.05,\,38)} = -1.686$. We will reject the null hypothesis if the calculated value of $t \leq -1.686$. If $t > -1.686$, we will not reject the null hypothesis. A left-tail rejection region is illustrated in Figure 3.3.
4. Using the food expenditure data, $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$. The value of the test statistic is
$$t = \frac{b_2 - 15}{\text{se}(b_2)} = \frac{10.21 - 15}{2.09} = -2.29$$
5. Since $t = -2.29 < -1.686$, we reject the null hypothesis that $\beta_2 \geq 15$ and accept the alternative that $\beta_2 < 15$. We conclude that households spend less than \$15 from each additional \$100 income on food.

EXAMPLE 3.5 | Two-Tail Test of an Economic Hypothesis

A consultant voices the opinion that based on other similar neighborhoods the households near the proposed market will spend an additional \$7.50 per additional \$100 income. In terms of our economic model, we can state this conjecture as the null hypothesis $\beta_2 = 7.5$. If we want to test whether this is true or not, then the alternative is that $\beta_2 \neq 7.5$. This alternative makes no claim about whether β_2 is greater than 7.5 or less than 7.5, simply that it is not 7.5. In such cases, we use a two-tail test, as follows:

1. The null hypothesis is $H_0: \beta_2 = 7.5$. The alternative hypothesis is $H_1: \beta_2 \neq 7.5$.
2. The test statistic $t = (b_2 - 7.5)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true.
3. Let us select $\alpha = 0.05$. The critical values for this two-tail test are the 2.5-percentile $t_{(0.025,\,38)} = -2.024$ and the 97.5-percentile $t_{(0.975,\,38)} = 2.024$. Thus, we will reject the null hypothesis if the calculated value of $t \geq 2.024$ **or** if $t \leq -2.024$. If $-2.024 < t < 2.024$, then we will not reject the null hypothesis.
4. For the food expenditure data, $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$. The value of the test statistic is
$$t = \frac{b_2 - 7.5}{\text{se}(b_2)} = \frac{10.21 - 7.5}{2.09} = 1.29$$
5. Since $-2.204 < t = 1.29 < 2.204$, we do not reject the null hypothesis that $\beta_2 = 7.5$. The sample data are consistent with the conjecture households will spend an additional \$7.50 per additional \$100 income on food.

We must avoid reading into this conclusion more than it means. We **do not** conclude from this test that $\beta_2 = 7.5$, only that the data are not incompatible with this parameter value. The data are also compatible with the null hypotheses $H_0: \beta_2 = 8.5$ ($t = 0.82$), $H_0: \beta_2 = 6.5$ ($t = 1.77$), and $H_0: \beta_2 = 12.5$ ($t = -1.09$). A hypothesis test **cannot** be used to prove that a null hypothesis is true.

There is a trick relating two-tail tests and confidence intervals that is sometimes useful. Let q be a value within a $100(1 - \alpha)\%$ confidence interval, so that if $t_c = t_{(1-\alpha/2,\,N-2)}$, then
$$b_k - t_c\text{se}(b_k) \leq q \leq b_k + t_c\text{se}(b_k)$$
If we test the null hypothesis $H_0: \beta_k = q$ against $H_1: \beta_k \neq q$, when q is inside the confidence interval, then we will *not* reject the null hypothesis at the level of significance α. If q is outside the confidence interval, then the two-tail test will reject the null hypothesis. We do not advocate using confidence intervals to test hypotheses, they serve a different purpose, but if you are given a confidence interval, this trick is handy.

EXAMPLE 3.6 | Two-Tail Test of Significance

While we are confident that a relationship exists between food expenditure and income, models are often proposed that are more speculative, and the purpose of hypothesis testing is to ascertain whether a relationship between variables exists or not. In this case, the null hypothesis is $\beta_2 = 0$; that is, no linear relationship exists between x and y. The alternative is $\beta_2 \neq 0$, which would mean that a relationship exists but that there may be either a positive or negative association between the variables. This is the most common form of a **test of significance**. The test steps are as follows:

1. The null hypothesis is $H_0: \beta_2 = 0$. The alternative hypothesis is $H_1: \beta_2 \neq 0$.
2. The test statistic $t = b_2/\text{se}(b_2) \sim t_{(N-2)}$ *if the null hypothesis is true*.
3. Let us select $\alpha = 0.05$. The critical values for this two-tail test are the 2.5-percentile $t_{(0.025,\,38)} = -2.024$ and the 97.5-percentile $t_{(0.975,\,38)} = 2.024$. We will reject the null hypothesis if the calculated value of $t \geq 2.024$ *or* if $t \leq -2.024$. If $-2.024 < t < 2.024$, we will not reject the null hypothesis.
4. Using the food expenditure data, $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$. The value of the test statistic is $t = b_2/\text{se}(b_2) = 10.21/2.09 = 4.88$.
5. Since $t = 4.88 > 2.024$, we reject the null hypothesis that $\beta_2 = 0$ and conclude that there is a statistically significant relationship between income and food expenditure.

Two points should be made about this result. First, the value of the t-statistic we computed in this two-tail test is the same as the value computed in the one-tail test of significance in Example 3.2. The difference between the two tests is the rejection region and the critical values. Second, the two-tail test of significance is something that should be done each time a regression model is estimated, and consequently, computer software automatically calculates the t-values for null hypotheses that the regression parameters are zero. Refer back to Figure 2.9. Consider the portion that reports the estimates:

Variable	Coefficient	Standard Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

Note that there is a column-labeled t-statistic. This is the t-statistic value for the null hypothesis that the corresponding parameter is zero. It is calculated as $t = b_k/\text{se}(b_k)$. Dividing the least squares estimates (Coefficient) by their standard errors (Std. error) gives the t-statistic values (t-statistic) for testing the hypothesis that the parameter is zero. The t-statistic value for the variable *INCOME* is 4.877381, which is relevant for testing the null hypothesis $H_0: \beta_2 = 0$. We have rounded this value to 4.88 in our discussions.

The t-value for testing the hypothesis that the intercept is zero equals 1.92. The $\alpha = 0.05$ critical values for these two-tail tests are $t_{(0.025,\,38)} = -2.024$ and $t_{(0.975,\,38)} = 2.024$ whether we are testing a hypothesis about the slope or intercept, so we fail to reject the null hypothesis that $H_0: \beta_1 = 0$ given the alternative $H_1: \beta_1 \neq 0$.

The final column, labeled "Prob.," is the subject of the following section.

Remark

"Statistically significant" does not necessarily imply "economically significant." For example, suppose that the CEO of a supermarket chain plans a certain course of action if $\beta_2 \neq 0$. Furthermore, suppose that a large sample is collected from which we obtain the estimate $b_2 = 0.0001$ with $\text{se}(b_2) = 0.00001$, yielding the t-statistic $t = 10.0$. We would reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 \neq 0$. Here, $b_2 = 0.0001$ is statistically different from zero. However, 0.0001 may not be "economically" different from zero, and the CEO may decide not to proceed with the plans. The message here is that one must think carefully about the importance of a statistical analysis before reporting or using the results.

3.5 The *p*-Value

When reporting the outcome of statistical hypothesis tests, it has become standard practice to report the ***p*-value** (an abbreviation for **probability value**) of the test. If we have the p-value of a

test, p, we can determine the outcome of the test by comparing the p-value to the chosen level of significance, α, *without* looking up or calculating the critical values. The rule is

p-Value Rule

Reject the null hypothesis when the p-value is less than, or equal to, the level of significance α. That is, if $p \leq \alpha$, then reject H_0. If $p > \alpha$, then do not reject H_0.

If you have chosen the level of significance to be $\alpha = 0.01, 0.05, 0.10$, or any other value, you can compare it to the p-value of a test and then reject, or not reject, without checking the critical value. In written works, reporting the p-value of a test allows the reader to apply his or her own judgment about the appropriate level of significance.

How the p-value is computed depends on the alternative. If t is the calculated value of the t-statistic, then

- if $H_1: \beta_k > c$, p = probability to the right of t
- if $H_1: \beta_k < c$, p = probability to the left of t
- if $H_1: \beta_k \neq c$, p = *sum* of probabilities to the right of $|t|$ *and* to the left of $-|t|$

Memory Trick

The direction of the alternative indicates the tail(s) of the distribution in which the p-value falls.

EXAMPLE 3.3 (continued) | *p*-Value for a Right-Tail Test

In Example 3.3, we tested the null hypothesis $H_0: \beta_2 \leq 5.5$ against the one-sided alternative $H_1: \beta_2 > 5.5$. The calculated value of the t-statistic was

$$t = \frac{b_2 - 5.5}{\text{se}(b_2)} = \frac{10.21 - 5.5}{2.09} = 2.25$$

In this case, since the alternative is "greater than" (>), the p-value of this test is the probability that a t-random variable with $N - 2 = 38$ degrees of freedom is greater than 2.25, or $p = P[t_{(38)} \geq 2.25] = 0.0152$.

This probability value cannot be found in the usual t-table of critical values, but it is easily found using the computer. Statistical software packages, and spreadsheets such as Excel, have simple commands to evaluate the *cumulative distribution function* (*cdf*) (see Appendix B.1) for a variety of probability distributions. If $F_X(x)$ is the *cdf* for a random variable X, then for any value $x = c$, the cumulative probability is $P[X \leq c] = F_X(c)$. Given such a function for the t-distribution, we compute the desired p-value as

$$p = P[t_{(38)} \geq 2.25] = 1 - P[t_{(38)} \leq 2.25] = 1 - 0.9848 = 0.0152$$

Following the p-value rule, we conclude that at $\alpha = 0.01$ we do not reject the null hypothesis. If we had chosen $\alpha = 0.05$, we would reject the null hypothesis in favor of the alternative.

The logic of the p-value rule is shown in Figure 3.5. The probability of obtaining a t-value greater than 2.25 is 0.0152, $p = P[t_{(38)} \geq 2.25] = 0.0152$. The 99th percentile $t_{(0.99,\,38)}$, which is the critical value for a right-tail test with the level of significance of $\alpha = 0.01$ must fall to the right of 2.25. This means that $t = 2.25$ does not fall in the rejection region if $\alpha = 0.01$ and we will not reject the null hypothesis at this level of significance. This is consistent with the *p-value rule*: When the p-value (0.0152) is greater than the chosen level of significance (0.01), we do not reject the null hypothesis.

On the other hand, the 95th percentile $t_{(0.95,\,38)}$, which is the critical value for a right-tail test with $\alpha = 0.05$, must be to the left of 2.25. This means that $t = 2.25$ falls in the rejection region, and we reject the null hypothesis at the level of significance $\alpha = 0.05$. This is consistent with the *p-value rule*: When the p-value (0.0152) is less than or equal to the chosen level of significance (0.05), we will reject the null hypothesis.

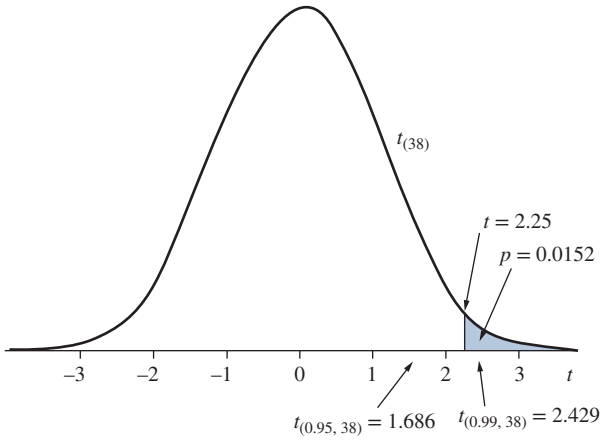


FIGURE 3.5 **The *p*-value for a right-tail test.**

EXAMPLE 3.4 (continued) | *p*-Value for a Left-Tail Test

In Example 3.4, we carried out a test with the rejection region in the left tail of the t-distribution. The null hypothesis was $H_0: \beta_2 \geq 15$, and the alternative hypothesis was $H_1: \beta_2 < 15$. The calculated value of the t-statistic was $t = -2.29$. To compute the p-value for this left-tail test, we calculate the probability of obtaining a t-statistic to the left of -2.29. Using your computer software, you will find this value to be $P[t_{(38)} \leq -2.29] = 0.0139$. Following the p-value rule, we conclude that at $\alpha = 0.01$, we do not reject the null hypothesis. If we choose $\alpha = 0.05$, we will reject the null hypothesis in favor of the alternative. See Figure 3.6 to see this graphically. Locate the 1st and 5th percentiles. These will be the critical values for left-tail tests with $\alpha = 0.01$ and $\alpha = 0.05$ levels of significance. When the p-value (0.0139) is greater than the level of significance ($\alpha = 0.01$), then the t-value -2.29 is not in the test rejection region. When the p-value (0.0139) is less than or equal to the level of significance ($\alpha = 0.05$), then the t-value -2.29 is in the test rejection region.

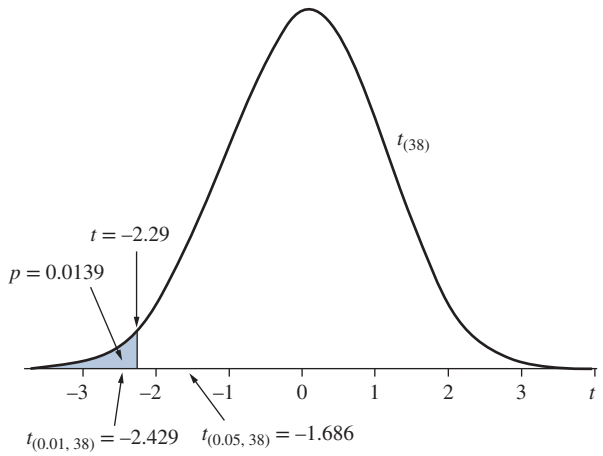


FIGURE 3.6 **The *p*-value for a left-tail test.**

EXAMPLE 3.5 (continued) | *p*-Value for a Two-Tail Test

For a two-tail test, the rejection region is in the two tails of the *t*-distribution, and the *p*-value is similarly calculated in the two tails of the distribution. In Example 3.5, we tested the null hypothesis that $\beta_2 = 7.5$ against the alternative hypothesis $\beta_2 \neq 7.5$. The calculated value of the *t*-statistic was $t = 1.29$. For this two-tail test, the *p*-value is the combined probability to the right of 1.29 and to the left of −1.29:

$$p = P\left[t_{(38)} \geq 1.29\right] + P\left[t_{(38)} \leq -1.29\right] = 0.2033$$

This calculation is depicted in Figure 3.7. Once the *p*-value is obtained, its use is unchanged. If we choose $\alpha = 0.05$, $\alpha = 0.10$, or even $\alpha = 0.20$, we will fail to reject the null hypothesis because $p > \alpha$.

At the beginning of this section, we stated the following rule for computing *p*-values for two-tail tests: if $H_1: \beta_k \neq c$, $p = sum$ of probabilities to the right of $|t|$ *and* to the left of $-|t|$. The reason for the use of absolute values in this rule is that it will apply equally well if the value of the *t*-statistic turns out to be positive or negative.

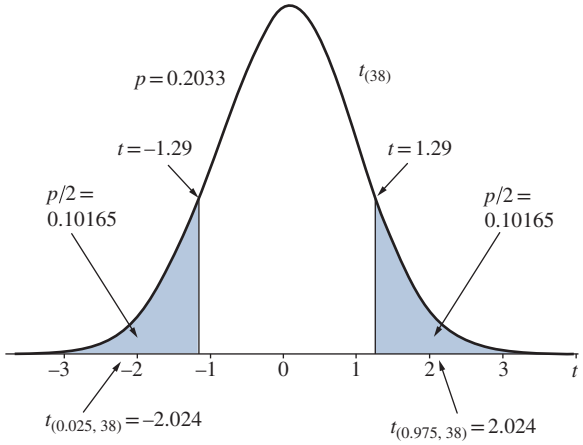


FIGURE 3.7 **The *p*-value for a two-tail test of significance.**

EXAMPLE 3.6 (continued) | *p*-Value for a Two-Tail Test of Significance

All statistical software computes the *p*-value for the two-tail test of significance for each coefficient when a regression analysis is performed. In Example 3.6, we discussed testing the null hypothesis $H_0: \beta_2 = 0$ against the alternative hypothesis $H_1: \beta_2 \neq 0$. For the calculated value of the *t*-statistic $t = 4.88$, the *p*-value is

$$p = P\left[t_{(38)} \geq 4.88\right] + P\left[t_{(38)} \leq -4.88\right] = 0.0000$$

Your software will automatically compute and report this *p*-value for a two-tail test of significance. Refer back to Figure 2.9 and consider just the portion reporting the estimates:

Variable	Coefficient	Standard Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

Next to each *t*-statistic value is the two-tail *p*-value, which is labeled "Prob." by the EViews software. Other software packages will use similar names. When inspecting computer output, we can immediately decide if an estimate is statistically significant (statistically different from zero using a two-tail test) by comparing the *p*-value to whatever level of significance we care to use. The estimated intercept has *p*-value 0.0622, so it is not statistically different from zero at the level of significance $\alpha = 0.05$, but it is statistically significant if $\alpha = 0.10$.

The estimated coefficient for income has a *p*-value that is zero to four places. Thus, $p \leq \alpha = 0.01$ or even $\alpha = 0.0001$, and thus, we reject the null hypothesis that income has no effect on food expenditure at these levels of significance. The *p*-value for this two-tail test of significance is not actually zero. If more places are used, then $p = 0.00001946$. Regression software usually does not print out more than four places because in practice levels of significance less than $\alpha = 0.001$ are rare.

3.6 Linear Combinations of Parameters

So far, we have discussed statistical inference (point estimation, interval estimation, and hypothesis testing) for a single parameter, β_1 or β_2. More generally, we may wish to estimate and test hypotheses about a **linear combination of parameters** $\lambda = c_1\beta_1 + c_2\beta_2$, where c_1 and c_2 are constants that we specify. One example is if we wish to estimate the expected value of a dependent

variable $E(y|x)$ when x takes some specific value, such as $x = x_0$. In this case, $c_1 = 1$ and $c_2 = x_0$, so that, $\lambda = c_1\beta_1 + c_2\beta_2 = \beta_1 + x_0\beta_2 = E(y|x = x_0)$.

Under assumptions SR1–SR5, the least squares estimators b_1 and b_2 are the best linear unbiased estimators of β_1 and β_2. It is also true that $\hat{\lambda} = c_1b_1 + c_2b_2$ is the best linear unbiased estimator of $\lambda = c_1\beta_1 + c_2\beta_2$. The estimator $\hat{\lambda}$ is unbiased because

$$E(\hat{\lambda}|\mathbf{x}) = E(c_1b_1 + c_2b_2|\mathbf{x}) = c_1E(b_1|\mathbf{x}) + c_2E(b_2|\mathbf{x}) = c_1\beta_1 + c_2\beta_2 = \lambda$$

Then, using the law of iterated expectations, $E(\hat{\lambda}) = E_{\mathbf{x}}\left[E(\hat{\lambda}|\mathbf{x})\right] = E_{\mathbf{x}}[\lambda] = \lambda$. To find the variance of $\hat{\lambda}$, recall from the Probability Primer, Section P.5.6, that if X and Y are random variables, and if a and b are constants, then the variance $\text{var}(aX + bY)$ is given in equation (P.20) as

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\,\text{cov}(X, Y)$$

In the estimator $(c_1b_1 + c_2b_2)$, both b_1 and b_2 are random variables, as we do not know what their values will be until a sample is drawn and estimates calculated. Applying (P.20), we have

$$\text{var}(\hat{\lambda}|\mathbf{x}) = \text{var}(c_1b_1 + c_2b_2|\mathbf{x}) = c_1^2\text{var}(b_1|\mathbf{x}) + c_2^2\text{var}(b_2|\mathbf{x}) + 2c_1c_2\text{cov}(b_1, b_2|\mathbf{x}) \tag{3.8}$$

The variances and covariances of the least squares estimators are given in (2.14)–(2.16). We estimate $\text{var}(\hat{\lambda}|\mathbf{x}) = \text{var}(c_1b_1 + c_2b_2|\mathbf{x})$ by replacing the unknown variances and covariances with their estimated variances and covariances in (2.20)–(2.22). Then

$$\widehat{\text{var}}(\hat{\lambda}|\mathbf{x}) = \widehat{\text{var}}(c_1b_1 + c_2b_2|\mathbf{x}) = c_1^2\widehat{\text{var}}(b_1|\mathbf{x}) + c_2^2\widehat{\text{var}}(b_2|\mathbf{x}) + 2c_1c_2\widehat{\text{cov}}(b_1, b_2|\mathbf{x}) \tag{3.9}$$

The standard error of $\hat{\lambda} = c_1b_1 + c_2b_2$ is the square root of the estimated variance,

$$\text{se}(\hat{\lambda}) = \text{se}(c_1b_1 + c_2b_2) = \sqrt{\widehat{\text{var}}(c_1b_1 + c_2b_2|\mathbf{x})} \tag{3.10}$$

If in addition SR6 holds, or if the sample is large, the least squares estimators b_1 and b_2 have normal distributions. It is also true that linear combinations of normally distributed variables are normally distributed, so that

$$\hat{\lambda}|\mathbf{x} = c_1b_1 + c_2b_2 \sim N\left[\lambda, \text{var}(\hat{\lambda}|\mathbf{x})\right]$$

where $\text{var}(\hat{\lambda}|\mathbf{x})$ is given in (3.8). You may be thinking of how long such calculations will take using a calculator, but don't worry. Most computer software will do the calculations for you. Now it's time for an example.

EXAMPLE 3.7 | Estimating Expected Food Expenditure

An executive might ask of the research staff, "Give me an estimate of average weekly food expenditure by households with $2,000 weekly income." Interpreting the executive's word "average" to mean "expected value," for the food expenditure model this means estimating

$$E(FOOD\_EXP|INCOME) = \beta_1 + \beta_2 INCOME$$

Recall that we measured income in $100 units in this example, so a weekly income of $2,000 corresponds to $INCOME = 20$. The executive is requesting an estimate of

$$E(FOOD\_EXP|INCOME = 20) = \beta_1 + \beta_2 20$$

which is a linear combination of the parameters.

Using the 40 observations in the data file *food*, in Section 2.3.2, we obtained the fitted regression,

$$\widehat{FOOD\_EXP} = 83.4160 + 10.2096 INCOME$$

The point estimate of average weekly food expenditure for a household with $2,000 income is

$$\widehat{E(FOOD\_EXP|INCOME = 20)} = b_1 + b_2 20$$
$$= 83.4160 + 10.2096(20) = 287.6089$$

We estimate that the expected food expenditure by a household with $2,000 income is $287.61 per week.

EXAMPLE 3.8 | An Interval Estimate of Expected Food Expenditure

If assumption SR6 holds, and given $\mathbf{x}$, the estimator $\hat{\lambda}$ has a normal distribution. We can form a standard normal random variable as

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\text{var}\left(\hat{\lambda}|\mathbf{x}\right)}} \sim N(0, 1)$$

Replacing the true variance in the denominator with the estimated variance, we form a pivotal t-statistic

$$t = \frac{\hat{\lambda} - \lambda}{\sqrt{\widehat{\text{var}}\left(\hat{\lambda}\right)}} = \frac{\hat{\lambda} - \lambda}{\text{se}\left(\hat{\lambda}\right)} = \frac{\left(c_1 b_1 + c_2 b_2\right) - \left(c_1\beta_1 + c_2\beta_2\right)}{\text{se}\left(c_1 b_1 + c_2 b_2\right)}$$

$$\sim t_{(N-2)} \tag{3.11}$$

If t_c is the $1 - \alpha/2$ percentile value from the $t_{(N-2)}$ distribution, then $P\left(-t_c \le t \le t_c\right) = 1 - \alpha$. Substitute (3.11) for t and rearrange to obtain

$$P\Big[\left(c_1 b_1 + c_2 b_2\right) - t_c \text{se}\left(c_1 b_1 + c_2 b_2\right) \le c_1\beta_1 + c_2\beta_2$$

$$\le \left(c_1 b_1 + c_2 b_2\right) + t_c \text{se}\left(c_1 b_1 + c_2 b_2\right)\Big] = 1 - \alpha$$

Thus, a $100(1 - \alpha)\%$ interval estimate for $c_1\beta_1 + c_2\beta_2$ is

$$\left(c_1 b_1 + c_2 b_2\right) \pm t_c \text{se}\left(c_1 b_1 + c_2 b_2\right)$$

In Example 2.5, we obtained the estimated covariance matrix

$$\begin{bmatrix} \widehat{\text{var}}(b_1) & \widehat{\text{cov}}(b_1, b_2) \\ \widehat{\text{cov}}(b_1, b_2) & \widehat{\text{var}}(b_2) \end{bmatrix} =$$

	C	*INCOME*
C	1884.442	−85.9032
INCOME	−85.9032	4.3818

To obtain the standard error for $b_1 + b_2 20$, we first calculate the estimated variance

$$\begin{aligned} \widehat{\text{var}}\left(b_1 + 20b_2\right) &= \widehat{\text{var}}\left(b_1\right) + \left(20^2 \times \widehat{\text{var}}\left(b_2\right)\right) \\ &\quad + \left(2 \times 20 \times \widehat{\text{cov}}\left(b_1, b_2\right)\right) \\ &= 1884.442 + \left(20^2 \times 4.3818\right) \\ &\quad + (2 \times 20 \times (-85.9032)) \\ &= 201.0169 \end{aligned}$$

Given $\widehat{\text{var}}\left(b_1 + 20b_2\right) = 201.0169$, the corresponding standard error is[2]

$$\begin{aligned} \text{se}\left(b_1 + 20b_2\right) &= \sqrt{\widehat{\text{var}}\left(b_1 + 20b_2\right)} = \sqrt{201.0169} \\ &= 14.1780 \end{aligned}$$

A 95% interval estimate of $E(FOOD\_EXP|INCOME = 20) = \beta_1 + \beta_2(20)$ is $\left(b_1 + b_2 20\right) \pm t_{(0.975,38)}\text{se}\left(b_1 + b_2 20\right)$ or

$$[287.6089 - 2.024(14.1780),\ 287.6089 + 2.024(14.1780)]$$

$$= [258.91,\ 316.31]$$

We estimate with 95% confidence that the expected food expenditure by a household with \$2,000 income is between \$258.91 and \$316.31.

3.6.1 Testing a Linear Combination of Parameters

So far, we have tested hypotheses involving only one regression parameter at a time. That is, our hypotheses have been of the form $H_0 : \beta_k = c$. A more **general linear hypothesis** involves both parameters and may be stated as

$$H_0 : c_1\beta_1 + c_2\beta_2 = c_0 \tag{3.12a}$$

where c_0, c_1, and c_2 are specified constants, with c_0 being the hypothesized value. Despite the fact that the null hypothesis involves both coefficients, it still represents a single hypothesis to be tested using a t-statistic. Sometimes, it is written equivalently in implicit form as

$$H_0 : \left(c_1\beta_1 + c_2\beta_2\right) - c_0 = 0 \tag{3.12b}$$

[2]The value 201.0169 was obtained using computer software. If you do the calculation by hand using the provided numbers, you obtain 201.034. Do not be alarmed if you obtain small differences like this occasionally, as it most likely is the difference between a computer-generated solution and a hand calculation.

The alternative hypothesis for the null hypothesis in (3.12a) might be

i. $H_1: c_1\beta_1 + c_2\beta_2 \neq c_0$ leading to a two-tail t-test
ii. $H_1: c_1\beta_1 + c_2\beta_2 > c_0$ leading to a right-tail t-test [Null may be "$\leq$"]
iii. $H_1: c_1\beta_1 + c_2\beta_2 < c_0$ leading to a left-tail t-test [Null may be "$\geq$"]

If the implicit form is used, the alternative hypothesis is adjusted as well.

The test of the hypothesis (3.12) uses the pivotal t-statistic

$$t = \frac{(c_1 b_1 + c_2 b_2) - c_0}{\text{se}(c_1 b_1 + c_2 b_2)} \sim t_{(N-2)} \text{ if the null hypothesis is true} \tag{3.13}$$

The rejection regions for the one- and two-tail alternatives (i)–(iii) are the same as those described in Section 3.3, and conclusions are interpreted the same way as well.

The form of the t-statistic is very similar to the original specification in (3.7). In the numerator, $(c_1 b_1 + c_2 b_2)$ is the best linear unbiased estimator of $(c_1\beta_1 + c_2\beta_2)$, and if the errors are normally distributed, or if we have a large sample, this estimator is normally distributed as well.

EXAMPLE 3.9 | Testing Expected Food Expenditure

The food expenditure model introduced in Section 2.1 and used as an illustration throughout provides an excellent example of how the **linear hypothesis** in (3.12) might be used in practice. For most medium and larger cities, there are forecasts of income growth for the coming year. A supermarket or food retail store of any type will consider this before a new facility is built. Their question is, if income in a locale is projected to grow at a certain rate, how much of that will be spent on food items? An executive might say, based on years of experience, "I expect that a household with $2,000 weekly income will spend, on average, more than $250 a week on food." How can we use econometrics to test this conjecture?

The regression function for the food expenditure model is

$$E(FOOD\_EXP|INCOME) = \beta_1 + \beta_2 INCOME$$

The executive's conjecture is that

$$E(FOOD\_EXP|INCOME = 20) = \beta_1 + \beta_2 20 > 250$$

To test the validity of this statement, we use it as the alternative hypothesis

$$H_1: \beta_1 + \beta_2 20 > 250, \text{ or } H_1: \beta_1 + \beta_2 20 - 250 > 0$$

The corresponding null hypothesis is the logical alternative to the executive's statement

$$H_0: \beta_1 + \beta_2 20 \leq 250, \text{ or } H_0: \beta_1 + \beta_2 20 - 250 \leq 0$$

Notice that the null and alternative hypotheses are in the same form as the general linear hypothesis with $c_1 = 1$, $c_2 = 20$, and $c_0 = 250$.

The rejection region for a right-tail test is illustrated in Figure 3.2. For a right-tail test at the $\alpha = 0.05$ level of significance, the t-critical value is the 95th percentile of the $t_{(38)}$ distribution, which is $t_{(0.95,\,38)} = 1.686$. If the calculated t-statistic value is greater than 1.686, we will reject the null hypothesis and accept the alternative hypothesis, which in this case is the executive's conjecture.

Computing the t-statistic value

$$t = \frac{(b_1 + 20b_2) - 250}{\text{se}(b_1 + 20b_2)}$$

$$= \frac{(83.4160 + 20 \times 10.2096) - 250}{14.1780}$$

$$= \frac{287.6089 - 250}{14.1780} = \frac{37.6089}{14.1780} = 2.65$$

Since $t = 2.65 > t_c = 1.686$, we reject the null hypothesis that a household with weekly income of $2,000 will spend $250 per week or less on food and conclude that the executive's conjecture that such households spend more than $250 is correct, with the probability of Type I error 0.05.

In Example 3.8, we estimated that a household with $2,000 weekly income will spend $287.6089, which is greater than the executive's speculated value of $250. However, simply observing that the estimated value is greater than $250 is not a statistical test. It might be numerically greater, but is it **significantly** greater? The t-test takes into account the precision with which we have estimated this expenditure level and also controls the probability of Type I error.

3.7 Exercises

3.7.1 Problems

3.1 There were 64 countries in 1992 that competed in the Olympics and won at least one medal. Let *MEDALS* be the total number of medals won, and let *GDPB* be GDP (billions of 1995 dollars). A linear regression model explaining the number of medals won is $MEDALS = \beta_1 + \beta_2 GDPB + e$. The estimated relationship is

$$\widehat{MEDALS} = b_1 + b_2 GDPB = 7.61733 + 0.01309 GDPB$$
$$\text{(se)} \qquad\qquad (2.38994)\ \ (0.00215) \qquad\qquad \text{(XR3.1)}$$

- **a.** We wish to test the hypothesis that there is no relationship between the number of medals won and *GDP* against the alternative there is a positive relationship. State the null and alternative hypotheses in terms of the model parameters.
- **b.** What is the test statistic for part (a) and what is its distribution if the null hypothesis is true?
- **c.** What happens to the distribution of the test statistic for part (a) if the alternative hypothesis is true? Is the distribution shifted to the left or right, relative to the usual *t*-distribution? [*Hint*: What is the expected value of b_2 if the null hypothesis is true, and what is it if the alternative is true?]
- **d.** For a test at the 1% level of significance, for what values of the *t*-statistic will we reject the null hypothesis in part (a)? For what values will we fail to reject the null hypothesis?
- **e.** Carry out the *t*-test for the null hypothesis in part (a) at the 1% level of significance. What is your economic conclusion? What does 1% level of significance mean in this example?

3.2 There were 64 countries in 1992 that competed in the Olympics and won at least one medal. Let *MEDALS* be the total number of medals won, and let *GDPB* be GDP (billions of 1995 dollars). A linear regression model explaining the number of medals won is $MEDALS = \beta_1 + \beta_2 GDPB + e$. The estimated relationship is given in equation (XR3.1) in Exercise 3.1.

- **a.** We wish to test the null hypothesis that a one-billion dollar increase in *GDP* leads to an increase in average, or expected, number of medals won by 0.015, against the alternative that it does not. State the null and alternative hypotheses in terms of the model parameters.
- **b.** What is the test statistic for part (a) and what is its distribution if the null hypothesis is true?
- **c.** What happens to the distribution of the test statistic in part (a) if the alternative hypothesis is true? Is the distribution shifted to the left or right, relative to the usual *t*-distribution, or is the direction of the shift uncertain? [*Hint*: What is the expected value of b_2 if the null hypothesis is true, and what is it if the alternative is true?]
- **d.** For a test at the 10% level of significance, for what values of the *t*-statistic, will we reject the null hypothesis in part (a)? For what values, will we fail to reject the null hypothesis?
- **e.** Carry out the *t*-test for the null hypothesis in part (a). What is your economic conclusion?
- **f.** If we carry out the test in part (a) at the 5% level of significance, what do we conclude? At the 1% level of significance, what do we conclude?
- **g.** Carry out the same test at the 5% level of significance, but changing the null hypothesis value of interest to 0.016, then 0.017. What is the calculated *t*-statistic value in each case? Which hypotheses do you reject, and which do you fail to reject?

3.3 There were 64 countries in 1992 that competed in the Olympics and won at least one medal. Let *MEDALS* be the total number of medals won, and let *GDPB* be GDP (billions of 1995 dollars). A linear regression model explaining the number of medals won is $MEDALS = \beta_1 + \beta_2 GDPB + e$. The estimated relationship is given in equation (XR3.1) in Exercise 3.1.

The estimated covariance between the slope and intercept estimators is −0.00181 and the estimated error variance is $\hat{\sigma}^2 = 320.336$. The sample mean of *GDPB* is $\overline{GDPB} = 390.89$ and the sample variance of *GDPB* is $s^2_{GDPB} = 1099615$.

- **a.** Estimate the expected number of medals won by a country with $GDPB = 25$.
- **b.** Calculate the standard error of the estimate in (a) using for the variance $\widehat{\text{var}}(b_1) + (25)^2 \widehat{\text{var}}(b_2) + (2)(25)\widehat{\text{cov}}(b_1, b_2)$.

c. Calculate the standard error of the estimate in (a) using for the variance $\hat{\sigma}^2 \left\{ (1/N) + \left[\left(25 - \overline{GDPB} \right)^2 \Big/ \left((N-1)\, s^2_{GDPB} \right) \right] \right\}$.

d. Construct a 95% interval estimate for the expected number of medals won by a country with $GDPB = 25$.

e. Construct a 95% interval estimate for the expected number of medals won by a country with $GDPB = 300$. Compare and contrast this interval estimate to that in part (d). Explain the differences you observe.

3.4 Assume that assumptions SR1–SR6 hold for the simple linear regression model, $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$. Generally, as the sample size N becomes larger, confidence intervals become narrower.

a. Is a narrower confidence interval for a parameter, such as β_2, desirable? Explain why or why not.

b. Give **two** specific reasons why, as the sample size gets larger, a confidence interval for β_2 tends to become narrower. The reasons should relate to the properties of the least squares estimator and/or interval estimation procedures.

3.5 If we have a large sample of data, then using critical values from the standard normal distribution for constructing a ***p*-value** is justified. But how large is "large"?

a. For a t-distribution with 30 degrees of freedom, the right-tail p-value for a t-statistic of 1.66 is 0.05366666. What is the approximate p-value using the cumulative distribution function of the standard normal distribution, $\Phi(z)$, in Statistical Table 1? Using a right-tail test with $\alpha = 0.05$, would you make the correct decision about the null hypothesis using the approximate p-value? Would the exact p-value be larger or smaller for a t-distribution with 90 degrees of freedom?

b. For a t-distribution with 200 degrees of freedom, the right-tail p-value for a t-statistic of 1.97 is 0.0251093. What is the approximate p-value using the standard normal distribution? Using a **two-tail test** with $\alpha = 0.05$, would you make the correct decision about the null hypothesis using the approximate p-value? Would the exact p-value be larger or smaller for a t-distribution with 90 degrees of freedom?

c. For a t-distribution with 1000 degrees of freedom, the right-tail p-value for a t-statistic of 2.58 is 0.00501087. What is the approximate p-value using the standard normal distribution? Using a two-tail test with $\alpha = 0.05$, would you make the correct decision about the null hypothesis using the approximate p-value? Would the exact p-value be larger or smaller for a t-distribution with 2000 degrees of freedom?

3.6 We have data on 2323 randomly selected households consisting of three persons in 2013. Let *ENTERT* denote the monthly entertainment expenditure ($) per person per month and let *INCOME* ($100) be monthly household income. Consider the simple linear regression model $ENTERT_i = \beta_1 + \beta_2 INCOME_i + e_i, i = 1, \ldots, 2323$. Assume that assumptions SR1–SR6 hold. The least squares estimated equation is $\widehat{ENTERT}_i = 9.820 + 0.503 INCOME_i$. The standard error of the slope coefficient estimator is $\text{se}(b_2) = 0.029$, the standard error of the intercept estimator is $\text{se}(b_1) = 2.419$, and the estimated covariance between the least squares estimators b_1 and b_2 is -0.062.

a. Construct a 90% confidence interval estimate for β_2 and interpret it for a group of CEOs from the entertainment industry.

b. The CEO of AMC Entertainment Mr. Lopez asks you to estimate the average monthly entertainment expenditure per person for a household with monthly income (for the three-person household) of $7500. What is your estimate?

c. AMC Entertainment's staff economist asks you for the estimated variance of the estimator $b_1 + 75b_2$. What is your estimate?

d. AMC Entertainment is planning to build a luxury theater in a neighborhood with average monthly income, for three-person households, of $7500. Their staff of economists has determined that in order for the theater to be profitable the average household will have to spend more than $45 per person per month on entertainment. Mr. Lopez asks you to provide conclusive statistical evidence, beyond reasonable doubt, that the proposed theater will be profitable. Carefully set up the null and alternative hypotheses, give the test statistic, and test rejection region using $\alpha = 0.01$. Using the information from the previous parts of the question, carry out the test and provide your result to the AMC Entertainment CEO.

e. The income elasticity of entertainment expenditures at the point of the means is $\varepsilon = \beta_2 \left(\overline{INCOME} \Big/ \overline{ENTERT} \right)$. The sample means of these variables are $\overline{ENTERT} = 45.93$ and

$\overline{INCOME} = 71.84$. Test the null hypothesis that the elasticity is 0.85 against the alternative that it is not 0.85, using the $\alpha = 0.05$ level of significance.

f. Using Statistical Table 1, compute the approximate two-tail p-value for the t-statistic in part (e). Using the p-value rule, do you reject the null hypothesis $\varepsilon = \beta_2\left(\overline{INCOME}/\overline{ENTERT}\right) = 0.85$, versus the alternative $\varepsilon \neq 0.85$, at the 10% level of significance? Explain.

3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME* on *BACHELOR* are

$$\begin{array}{lccc} \widehat{INCOME} = & (a) & + & 1.029BACHELOR \\ \text{se} & (2.672) & & (c) \\ t & (4.31) & & (10.75) \end{array}$$

a. Using the information provided calculate the estimated intercept. Show your work.

b. Sketch the estimated relationship. Is it increasing or decreasing? Is it a positive or inverse relationship? Is it increasing or decreasing at a constant rate or is it increasing or decreasing at an increasing rate?

c. Using the information provided calculate the standard error of the slope coefficient. Show your work.

d. What is the value of the t-statistic for the null hypothesis that the intercept parameter equals 10?

e. The p-value for a two-tail test that the intercept parameter equals 10, from part (d), is 0.572. Show the p-value in a sketch. On the sketch, show the rejection region if $\alpha = 0.05$.

f. Construct a 99% interval estimate of the slope. Interpret the interval estimate.

g. Test the null hypothesis that the slope coefficient is one against the alternative that it is not one at the 5% level of significance. State the economic result of the test, in the context of this problem.

3.8 Using 2011 data on 141 U.S. public research universities, we examine the relationship between cost per student and full-time university enrollment. Let *ACA* = real academic cost per student (thousands of dollars), and let *FTESTU* = full-time student enrollment (thousands of students). The least squares fitted relation is $\widehat{ACA} = 14.656 + 0.266FTESTU$.

a. For the regression, the 95% interval estimate for the intercept is [10.602, 18.710]. Calculate the standard error of the estimated intercept.

b. From the regression output, the standard error for the slope coefficient is 0.081. Test the null hypothesis that the true slope, β_2, is 0.25 (or less) against the alternative that the true slope is greater than 0.25 using the 10% level of significance. Show all steps of this hypothesis test, including the null and alternative hypotheses, and state your conclusion.

c. On the regression output, the automatically provided p-value for the estimated slope is 0.001. What is the meaning of this value? Use a sketch to illustrate your answer.

d. A member of the board of supervisors states that *ACA* should fall if we admit more students. Using the estimated equation and the information in parts (a)–(c), test the null hypothesis that the slope parameter β_2 is zero, or positive, against the alternative hypothesis that it is negative. Use the 5% level of significance. Show all steps of this hypothesis test, including the null and alternative hypotheses, and state your conclusion. Is there any statistical support for the board member's conjecture?

e. In 2011, Louisiana State University (LSU) had a full-time student enrollment of 27,950. Based on the estimated equation, the least squares estimate of $E(ACA|FTESTU = 27,950)$ is 22.079, with standard error 0.964. The actual value of *ACA* for LSU that year was 21.403. Would you say that this value is surprising or not surprising? Explain.

3.9 Using data from 2013 on 64 black females, the estimated linear regression between *WAGE* (earnings per hour, in $) and years of education, *EDUC* is $\widehat{WAGE} = -8.45 + 1.99EDUC$.

a. The standard error of the estimated slope coefficient is 0.52. Construct and interpret a 95% interval estimate for the effect of an additional year of education on a black female's expected hourly wage rate.

b. The standard error of the estimated intercept is 7.39. Test the null hypothesis that the intercept $\beta_1 = 0$ against the alternative that the true intercept is not zero, using the $\alpha = 0.10$ level of significance. In your answer, show (i) the formal null and alternative hypotheses, (ii) the test statistic and

its distribution under the null hypothesis, (iii) the rejection region (in a figure), (iv) the calculated value of the test statistic, and (v) state your conclusion, with its economic interpretation.

c. Estimate the expected wage for a black female with 16 years of education, $E(WAGE|EDUC = 16)$.

d. The estimated covariance between the intercept and slope is −3.75. Construct a 95% interval estimate for the expected wage for a black female with 16 years of education.

e. It is conjectured that a black female with 16 years of education will have an expected wage of more than $23 per hour. Use this as the "alternative hypothesis" in a test of the conjecture at the 10% level of significance. Does the evidence support the conjecture or not?

3.10 Using data from 2013 on 64 black females, the estimated log-linear regression between *WAGE* (earnings per hour, in $) and years of education, *EDUC* is $\widehat{\ln(WAGE)} = 1.58 + 0.09EDUC$. The reported t-statistic for the slope coefficient is 3.95.

a. Test at the 5% level of significance, the null hypothesis that the return to an additional year of education is less than or equal to 8% against the alternative that the rate of return to education is more than 8%. In your answer, show (i) the formal null and alternative hypotheses, (ii) the test statistic and its distribution under the null hypothesis, (iii) the rejection region (in a figure), (iv) the calculated value of the test statistic, and (v) state your conclusion, with its economic interpretation.

b. Testing the null hypothesis that the return to education is 8%, against the alternative that it is not 8%, we obtain the p-value 0.684. What is the p-value for the test in part (a)? In a sketch, show for the test in part (a) the p-value and the 5% critical value from the t-distribution.

c. Construct a 90% interval estimate for the return to an additional year of education and state its interpretation.

3.11 The theory of labor supply indicates that more labor services will be offered at higher wages. Suppose that *HRSWK* is the usual number of hours worked per week by a randomly selected person and *WAGE* is their hourly wage. Our regression model is specified as $HRSWK = \beta_1 + \beta_2 WAGE + e$. Using a sample of 9799 individuals from 2013, we obtain the estimated regression $\widehat{HRSWK} = 41.58 + 0.011WAGE$. The estimated variances and covariance of the least squares estimators are as follows:

	INTERCEPT	*WAGE*
INTERCEPT	0.02324	−0.00067
WAGE	−0.00067	0.00003

a. Test the null hypothesis that the relationship has slope that is less than, or equal to, zero at the 5% level of significance. State the null and alternative hypotheses in terms of the model parameters. Using the results, do we confirm or refute the theory of labor supply?

b. Use Statistical Table 1 of normal probabilities to calculate an approximate p-value for the test in (a). Draw a sketch representing the p-value.

c. Under assumptions SR1–SR6 of the simple regression model, the expected number of hours worked per week is $E(HRSWK|WAGE) = \beta_1 + \beta_2 WAGE$. Construct a 95% interval estimate for the expected number of hours worked per week for a person earning $20/h.

d. In the sample, there are 203 individuals with hourly wage $20. The average number of hours worked for these people is 41.68. Is this result compatible with the interval estimate in (c)? Explain your reasoning.

e. Test the null hypothesis that the expected hours worked for a person earning $20 per hour is 41.68, against the alternative that it is not, at the 1% level of significance.

3.12 Consider a log-linear regression for the weekly sales (number of cans) of a national brand of canned tuna (*SAL*1 = target brand sales) as a function of the ratio of its price to the price of a competitor, *RPRICE*3 = 100(price of target brand ÷ price competitive brand #3), $\ln(SAL1) = \gamma_1 + \gamma_2 RPRICE3 + e$. Using $N = 52$ weekly observations the least squares estimated equation is

$$\ln(SAL1) = 11.481 - 0.031RPRICE3$$
$$\text{(se)} \qquad (0.535) \quad (0.00529)$$

a. The variable *RPRICE*3 is the price of the target brand as a percentage of the price of competitive brand #3 or more simply "the relative price." The sample mean of *RPRICE*3 is 99.66, its median

is 100, its minimum value is 70.11, and its maximum value is 154.24. What do these summary statistics tell us about the prices of the target brand relative to the prices of its competitor?

b. Interpret the coefficient of *RPRICE*3. Does its sign make economic sense?

c. Construct and interpret a 95% interval estimate for the effect on the weekly sales, *SAL*1, of a 1% increase in the price of the target brand as a percentage of the price of competitive brand #3, which is relative price *RPRICE*3.

d. Carry out a test of the null hypothesis $H_0: \gamma_2 \geq -0.02$ against the alternative $H_1: \gamma_2 < -0.02$ using the $\alpha = 0.01$ level of significance. Include in your answer (i) the test statistic and its distribution if the null hypothesis is true, (ii) a sketch of the rejection region, (iii) show the location of the test statistic value, (iv) state your conclusion, and (v) show on the sketch the region that would represent the *p*-value.

e. "Hypothesis tests and interval estimators for the regression model are valid as long as the regression error terms are normally distributed." Is this true or false? Explain.

3.13 Consider the following estimated area response model for sugar cane (area of sugar cane planted in thousands of hectares in a region of Bangladesh), as a function of relative price (100 times the price of sugar cane divided by the price of jute, which is an alternative crop to sugar cane, planted by Bangladesh farmers), $\widehat{AREA}_t = -0.24 + 0.50RPRICE_t$ using 34 annual observations.

a. The sample average of *RPRICE* is 114.03, with a minimum of 74.9 and a maximum of 182.2. *RPRICE* is the price of sugar cane taken as a percentage of the price of jute. What do these sample statistics tell us about the relative price of sugar cane?

b. Interpret the intercept and slope of the estimated relation.

c. The *t*-statistic is −0.01 for the hypothesis that the intercept parameter is zero. What do you conclude? Is this an economically surprising result? Explain.

d. The sample mean area planted is 56.83 thousand hectares, and the sample mean for relative price is 114.03. Taking these values as given, test at the 5% level of significance the hypothesis that the elasticity of area response to price at the means is 1.0. The estimated variance of the coefficient of *RPRICE* is 0.020346.

e. The model is re-estimated in log-linear form, obtaining $\widehat{\ln(AREA_t)} = 3.21 + 0.0068RPRICE_t$. Interpret the coefficient of *RPRICE*. The standard error of the slope estimate is 0.00229. What does that tell us about the estimated relationship?

f. Using the model in (e), test the null hypothesis that a 1% increase in the price of sugar cane relative to the price of jute increases the area planted in sugar cane by 1%. Use the 5% level of significance and a two-tail test. Include (i) the test statistic and its distribution if the null hypothesis is true, (ii) a sketch of the rejection region, (iii) show the location of the test statistic value, (iv) state your conclusion, and (v) show on the sketch, the region that would represent the *p*-value.

3.14 What is the meaning of statistical significance and how valuable is this concept? A *t*-statistic is $t = (b - c)/\text{se}(b)$, where b is an estimate of a parameter β, c is the hypothesized value, and $\text{se}(b)$ is the standard error. If the sample size N is large, then the statistic is approximately a standard normal distribution if the null hypothesis $\beta = c$ is true.

a. With a 5% level of significance, we assert that an event happening with less than a one in 20 chance is "statistically significant," while an event happening with more than a one in 20 chance is not statistically significant. True or False?

b. Would you say something happening one time in 10 by chance (10%) is very improbable or not very improbable? Would you say something happening one time in 100 by chance (1%) is very improbable or not?

c. If we adopt a rule that in large samples, a *t*-value greater than 2.0 (in absolute value) indicates statistical significance, and we use Statistical Table 1 of standard normal cumulative probabilities, what is the implied significance level? If we adopt a rule that in large samples, a *t*-value greater than 3.0 (in absolute value) indicates statistical significance, what is the implied significance level?

d. Suppose that we clinically test two diet pills, one called "Reliable" and another called "More." Using the Reliable pill, the estimated weight loss is 5 lbs with a standard error of 0.5 lbs. With the More pill, the estimated weight loss is 20 lbs with standard error 10 lbs. When testing whether the true weight loss is zero (the null, or none, hypothesis), what are the *t*-statistic values? What is the ratio of the *t*-values?

e. If the drugs Reliable and More were equivalent in safety, cost and every other comparison, and if your goal was weight loss, which drug would you take? Why?

3.15 In a capital murder trial, with a potential penalty of life in prison, would you as judge tell the jury to make sure that we accidently convict an innocent person only one time in a hundred, or use some other threshold? What would it be?

a. What is the economic cost of a Type I error in this example? List some of the factors that would have to be considered in such a calculation.

b. What is the economic cost of a Type II error in this example? List some of the factors that would have to be considered in such a calculation.

3.16 A big question in the United States, a question of "cause and effect," is whether mandatory health care will really make Americans healthier. What is the role of hypothesis testing in such an investigation?

a. Formulate null and alternative hypotheses based on the question.

b. What is a Type I error in the context of this question? What factors would you consider if you were assigned the task of calculating the economic cost of a Type I error in this example?

c. What is a Type II error in the context of this question? What factors would you consider if you were assigned the task of calculating the economic cost of a Type II error in this example?

d. If we observe that individuals who have health insurance are in fact healthier, does this prove that we should have mandatory health care?

e. There is a saying, "Correlation does not imply causation." How might this saying relate to part (d)?

f. Post hoc ergo propter hoc (Latin: "after this, therefore because of this") is a logical fallacy discussed widely in Principles of Economics textbooks. An example might be "A rooster crows and then the sun appears, thus the crowing rooster causes the sun to rise." How might this fallacy relate to the observation in part (d)?

3.17 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$. Where *WAGE* is hourly wage rate in US 2013 dollars. *EDUC* is years of schooling. The model is estimated twice, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \qquad \widehat{WAGE} = -10.76 + 2.46EDUC, \quad N = 986$$
$$\text{(se)} \qquad (2.27) \quad (0.16)$$

$$\text{Rural} \qquad \widehat{WAGE} = -4.88 + 1.80EDUC, \quad N = 214$$
$$\text{(se)} \qquad (3.29) \quad (0.24)$$

a. Using the urban regression, test the null hypothesis that the regression slope equals 1.80 against the alternative that it is greater than 1.80. Use the $\alpha = 0.05$ level of significance. Show all steps, including a graph of the critical region and state your conclusion.

b. Using the rural regression, compute a 95% interval estimate for expected *WAGE* if $EDUC = 16$. The required standard error is 0.833. Show how it is calculated using the fact that the estimated covariance between the intercept and slope coefficients is -0.761.

c. Using the urban regression, compute a 95% interval estimate for expected *WAGE* if $EDUC = 16$. The estimated covariance between the intercept and slope coefficients is -0.345. Is the interval estimate for the urban regression wider or narrower than that for the rural regression in (b). Do you find this plausible? Explain.

d. Using the rural regression, test the hypothesis that the intercept parameter β_1 equals four, or more, against the alternative that it is less than four, at the 1% level of significance.

3.18 A life insurance company examines the relationship between the amount of life insurance held by a household and household income. Let *INCOME* be household income (thousands of dollars) and *INSURANCE* the amount of life insurance held (thousands of dollars). Using a random sample of $N = 20$ households, the least squares estimated relationship is

$$\widehat{INSURANCE} = 6.855 + 3.880INCOME$$
$$\text{(se)} \qquad (7.383) \quad (0.112)$$

a. Draw a sketch of the fitted relationship identifying the estimated slope and intercept. The sample mean of $INCOME = 59.3$. What is the sample mean of the amount of insurance held? Locate the point of the means in your sketch.

b. How much do we estimate that the average amount of insurance held changes with each additional \$1000 of household income? Provide both a point estimate and a 95% interval estimate. Explain the interval estimate to a group of stockholders in the insurance company.

c. Construct a 99% interval estimate of the expected amount of insurance held by a household with \$100,000 income. The estimated covariance between the intercept and slope coefficient is −0.746.

d. One member of the management board claims that for every \$1000 increase in income the average amount of life insurance held will increase by \$5000. Let the algebraic model be $INSURANCE = \beta_1 + \beta_2 INCOME + e$. Test the hypothesis that the statement is true against the alternative that it is not true. State the conjecture in terms of a null and alternative hypothesis about the model parameters. Use the 5% level of significance. Do the data support the claim or not? Clearly, indicate the test statistic used and the rejection region.

e. Test the hypothesis that as income increases the amount of life insurance held increases by the same amount. That is, test the null hypothesis that the slope is one. Use as the alternative that the slope is larger than one. State the null and alternative hypotheses in terms of the model parameters. Carry out the test at the 1% level of significance. Clearly indicate the test statistic used, and the rejection region. What is your conclusion?

3.7.2 Computer Exercises

3.19 The owners of a motel discovered that a defective product was used during construction. It took 7 months to correct the defects during which approximately 14 rooms in the 100-unit motel were taken out of service for 1 month at a time. The data are in the file *motel*.

a. Plot *MOTEL_PCT* and *COMP_PCT* versus *TIME* on the same graph. What can you say about the occupancy rates over time? Do they tend to move together? Which seems to have the higher occupancy rates? Estimate the regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$. Construct a 95% interval estimate for the parameter β_2. Have we estimated the association between *MOTEL_PCT* and *COMP_PCT* relatively precisely, or not? Explain your reasoning.

b. Construct a 90% interval estimate of the expected occupancy rate of the motel in question, *MOTEL_PCT*, given that $COMP\_PCT = 70$.

c. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0 : \beta_2 \leq 0$ against the alternative hypothesis $H_0 : \beta_2 > 0$ at the $\alpha = 0.01$ level of significance. Discuss your conclusion. Clearly define the test statistic used and the rejection region.

d. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0 : \beta_2 = 1$ against the alternative hypothesis $H_0 : \beta_2 \neq 1$ at the $\alpha = 0.01$ level of significance. If the null hypothesis were true, what would that imply about the motel's occupancy rate versus their competitor's occupancy rate? Discuss your conclusion. Clearly define the test statistic used and the rejection region.

e. Calculate the least squares residuals from the regression of *MOTEL_PCT* on *COMP_PCT* and plot them against *TIME*. Are there any unusual features to the plot? What is the predominant sign of the residuals during time periods 17–23 (July, 2004 to January, 2005)?

3.20 The owners of a motel discovered that a defective product was used during construction. It took seven months to correct the defects during which approximately 14 rooms in the 100-unit motel were taken out of service for one month at a time. The data are in the file *motel*.

a. Calculate the sample average occupancy rate for the motel during the time when there were no repairs being made. What is the sample average occupancy rate for the motel during the time when there were repairs being made? How big a difference is there?

b. Consider the linear regression $MOTEL\_PCT = \delta_1 + \delta_2 REPAIR + e$, where *REPAIR* is an indicator variable taking the value 1 during the repair period and 0 otherwise. What are the estimated coefficients? How do these estimated coefficients relate to the calculations in part (a)?

c. Construct a 95% interval estimate for the parameter δ_2 and give its interpretation. Have we estimated the effect of the repairs on motel occupancy relatively precisely, or not? Explain.

d. The motel wishes to claim economic damages because the faulty materials led to repairs which cost them customers. To do so, their economic consultant tests the null hypothesis $H_0 : \delta_2 \geq 0$ against the alternative hypothesis $H_1 : \delta_2 < 0$. Explain the logic behind stating the null and alternative hypotheses in this way. Carry out the test at the $\alpha = 0.05$ level of significance. Discuss your conclusions. Clearly state the test statistic, the rejection region, and the *p*-value.

e. To further the motel's claim, the consulting economist estimates a regression model $(MOTEL\_PCT - COMP\_PCT) = \gamma_1 + \gamma_2 REPAIR + e$, so that the dependent variable is the difference in the occupancy rates. Construct and discuss the economic meaning of the 95% interval estimate of γ_2.

f. Test the null hypothesis that $\gamma_2 = 0$ against the alternative that $\gamma_2 < 0$ at the $\alpha = 0.01$ level of significance. Discuss the meaning of the test outcome. Clearly state the test statistic, the rejection region, and the *p*-value.

3.21 The capital asset pricing model (CAPM) is described in Exercise 2.16. Use all available observations in the data file *capm5* for this exercise.

a. Construct 95% interval estimates of Exxon-Mobil's and Microsoft's "beta." Assume that you are a stockbroker. Explain these results to an investor who has come to you for advice.

b. Test at the 5% level of significance the hypothesis that Ford's "beta" value is one against the alternative that it is not equal to one. What is the economic interpretation of a beta equal to one? Repeat the test and state your conclusions for General Electric's stock and Exxon-Mobil's stock. Clearly state the test statistic used and the rejection region for each test, and compute the *p*-value.

c. Test at the 5% level of significance the null hypothesis that Exxon-Mobil's "beta" value is greater than or equal to one against the alternative that it is less than one. Clearly state the test statistic used and the rejection region for each test, and compute the *p*-value. What is the economic interpretation of a beta less than one?

d. Test at the 5% level of significance the null hypothesis that Microsoft's "beta" value is less than or equal to one against the alternative that it is greater than one. Clearly state the test statistic used and the rejection region for each test, and compute the *p*-value. What is the economic interpretation of a beta more than one?

e. Test at the 5% significance level, the null hypothesis that the intercept term in the CAPM model for Ford's stock is zero, against the alternative that it is not. What do you conclude? Repeat the test and state your conclusions for General Electric's stock and Exxon-Mobil's stock. Clearly state the test statistic used and the rejection region for each test, and compute the *p*-value.

3.22 The data file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in $1000 units), *PRICE*, and total interior area in hundreds of square feet, *SQFT*.

a. Using the linear regression $PRICE = \beta_1 + \beta_2 SQFT + e$, estimate the elasticity of expected house *PRICE* with respect to *SQFT*, evaluated at the sample means. Construct a 95% interval estimate for the elasticity, treating the sample means as if they are given (not random) numbers. What is the interpretation of the interval?

b. Test the null hypothesis that the elasticity, calculated in part (a), is one against the alternative that the elasticity is not one. Use the 1% level of significance. Clearly state the test statistic used, the rejection region, and the test *p*-value. What do you conclude?

c. Using the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$, test the hypothesis that the marginal effect on expected house price of increasing house size by 100 square feet is less than or equal to $13000 against the alternative that the marginal effect will be greater than $13000. Use the 5% level of significance. Clearly state the test statistic used, the rejection region, and the test *p*-value. What do you conclude?

d. Using the linear regression $PRICE = \beta_1 + \beta_2 SQFT + e$, estimate the expected price, $E(PRICE|SQFT) = \beta_1 + \beta_2 SOFT$, for a house of 2000 square feet. Construct a 95% interval estimate of the expected price. Describe your interval estimate to a general audience.

e. Locate houses in the sample with 2000 square feet of living area. Calculate the sample mean (average) of their selling prices. Is the sample average of the selling price for houses with $SQFT = 20$ compatible with the result in part (d)? Explain.

3.23 The data file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price in $1000 units, *PRICE*, and total interior area in hundreds of square feet, *SQFT*.

a. Using the quadratic regression model, $PRICE = \alpha_1 + \alpha_2 SOFT^2 + e$, test the hypothesis that the marginal effect on expected house price of increasing the size of a 2000 square foot house by 100 square feet is less than or equal to $13000 against the alternative that the marginal effect will be greater than $13000. Use the 5% level of significance. Clearly state the test statistic used, the rejection region, and the test *p*-value. What do you conclude?

b. Using the quadratic regression model in part (a), test the hypothesis that the marginal effect on expected house price of increasing the size of a 4000 square foot house by 100 square feet is less than or equal to $13000 against the alternative that the marginal effect will be greater than $13000.

Use the 5% level of significance. Clearly state the test statistic used, the rejection region, and the test p-value. What do you conclude?

c. Using the quadratic regression model in part (a), estimate the expected price $E(PRICE|SQFT) = \alpha_1 + \alpha_2 SQFT^2$ for a house of 2000 square feet. Construct a 95% interval estimate of the expected price. Describe your interval estimate to a general audience.

d. Locate houses in the sample with 2000 square feet of living area. Calculate the sample mean (average) of their selling prices. Is the sample average of the selling price for houses with $SQFT = 20$ compatible with the result in part (c)? Explain.

3.24 We introduced Professor Ray C. Fair's model for explaining and predicting U.S. presidential elections in Exercise 2.23. Fair's data, 26 observations for the election years from 1916 to 2016, are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic party. Define *GROWTH* = *INCUMB* × *growth rate*, where growth rate is the annual rate of change in real per capita GDP in the first three quarters of the election year. If Democrats are the incumbent party, then *INCUMB* = 1; if the Republicans are the incumbent party then *INCUMB* = −1.

a. Estimate the linear regression, $VOTE = \beta_1 + \beta_2 GROWTH + e$, using data from 1916 to 2016. Construct a 95% interval estimate of the effect of economic growth on expected *VOTE*. How would you describe your finding to a general audience?

b. The expected *VOTE* in favor of the Democratic candidate is $E(VOTE|GROWTH) = \beta_1 + \beta_2 GROWTH$. Estimate $E(VOTE|GROWTH = 4)$ and construct a 95% interval estimate and a 99% interval estimate. Assume a Democratic incumbent is a candidate for a second presidential term. Is achieving a 4% growth rate enough to ensure a victory? Explain.

c. Test the hypothesis that when *INCUMB* = 1 economic growth has either a zero or negative effect on expected *VOTE* against the alternative that economic growth has a positive effect on expected *VOTE*. Use the 1% level of significance. Clearly state the test statistic used, the rejection region, and the test p-value. What do you conclude?

d. Define *INFLAT* = *INCUMB* × inflation rate, where the inflation rate is the growth in prices over the first 15 quarters of an administration. Using the data from 1916 to 2016, and the model $VOTE = \alpha_1 + \alpha_2 INFLAT + e$, test the hypothesis that inflation has no effect against the alternative that it does have an effect. Use the 1% level of significance. State the test statistic used, the rejection region, and the test p-value and state your conclusion.

3.25 Using data on the "Ashcan School," we have an opportunity to study the market for art. What factors determine the value of a work of art? Use the data in the file *ashcan_small*. [Note: the file *ashcan* contains more variables.]

a. Define *YEARS_OLD* = *DATE_AUCTN* − *CREATION*, which is the age of the painting at the time of its sale. Use data on works that sold (*SOLD* = 1) to estimate the regression $\ln(RHAMMER) = \beta_1 + \beta_2 YEARS\_OLD + e$. Construct a 95% interval estimate for the percentage change in real hammer price given that a work of art is another year old at the time of sale. [*Hint*: Review the discussion of equation (2.28).] Explain the result to a potential art buyer.

b. Test the null hypothesis that each additional year of age increases the "hammer price" by 2%, against the two-sided alternative. Use the 5% level of significance.

c. The variable *DREC* is an indicator variable taking the value one if a sale occurred during a recession and is zero otherwise. Use data on works that sold (*SOLD* = 1) to estimate the regression model $\ln(RHAMMER) = \alpha_1 + \alpha_2 DREC + e$. Construct a 95% interval estimate of the percentage reduction in hammer price when selling in a recession. Explain your finding to a client who is considering selling during a recessionary period.

d. Test the conjecture that selling a work of art during a recession reduces the hammer price by 2% or less, against the alternative that the reduction in hammer price is greater than 2%. Use the 5% level of significance. Clearly state the test statistic used, the rejection region, and the test p-value. What is your conclusion?

3.26 How much does experience affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, experience, and other variables from the March 2013 Current Population Survey (CPS). [Note: The data file *cps5* contains more observations and variables.]

a. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER + e$ and discuss the results.

b. Test the statistical significance of the estimated relationship at the 5% level. Use a one-tail test. What is your alternative hypothesis? What do you conclude?

c. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER + e$ for individuals living in a metropolitan area, where $METRO = 1$. Is there a statistically significant positive relationship between expected wages and experience at the 1% level? How much of an effect is there?

d. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER + e$ for individuals not living in a metropolitan area, where $METRO = 0$. Is there a statistically significant positive relationship between expected wages and experience at the 1% level? Can we safely say that experience has no effect on wages for individuals living in nonmetropolitan areas? Explain.

3.27 Is the relationship between experience and wages constant over one's lifetime? We will investigate this question using a quadratic model. The data file *cps5_small* contains 1200 observations on hourly wage rates, experience, and other variables from the March 2013 Current Population Survey (CPS). [Note: the data file *cps5* contains more observations and variables.]

a. Create the variable $EXPER30 = EXPER - 30$. Describe this variable. When is it positive, negative or zero?

b. Estimate by least squares the quadratic model $WAGE = \gamma_1 + \gamma_2 (EXPER30)^2 + e$. Test the null hypothesis that $\gamma_2 = 0$ against the alternative $\gamma_2 \neq 0$ at the 1% level of significance. Is there a statistically significant quadratic relationship between expected *WAGE* and *EXPER*30?

c. Create a plot of the fitted value $\widehat{WAGE} = \hat{\gamma}_1 + \hat{\gamma}_2 (EXPER30)^2$, on the *y*-axis, versus *EXPER*30, on the *x*-axis. Up to the value $EXPER30 = 0$ is the slope of the plot constant, or is it increasing, or decreasing? Up to the value $EXPER30 = 0$ is the function increasing at an increasing rate or increasing at a decreasing rate?

d. If $y = a + bx^2$ then $dy/dx = 2bx$. Using this result, calculate the estimated slope of the fitted function $\widehat{WAGE} = \hat{\gamma}_1 + \hat{\gamma}_2 (EXPER30)^2$, when $EXPER = 0$, when $EXPER = 10$, and when $EXPER = 20$.

e. Calculate the *t*-statistic for the null hypothesis that the slope of the function is zero, $H_0 : 2\gamma_2 EXPER30 = 0$, when $EXPER = 0$, when $EXPER = 10$, and when $EXPER = 20$.

3.28 The owners of a motel discovered that a defective product was used during construction. It took 7 months to correct the defects during which approximately 14 rooms in the 100-unit motel were taken out of service for 1 month at a time. The data are in the file *motel*.

a. Create a new variable, $RELPRICE2 = 100 RELPRICE$, which equals the percentage of the competitor's price charged by the motel in question. Plot *RELPRICE*2 against *TIME*. Compute the summary statistics for this variable. What are the sample mean and median? What are the minimum and maximum values? Does the motel in question charge more than its competitors for a room, or less, or about the same? Explain.

b. Consider a linear regression with $y = MOTEL\_PCT$ and $x = RELPRICE2$. Interpret the estimated slope coefficient. Construct a 95% interval estimate for the slope. Have we estimated the slope of the relationship very well? Explain your answer.

c. Construct a 90% interval estimate of the expected motel occupancy rate if the motel's price is 80% of its competitor's price. Do you consider the interval relatively narrow or relatively wide? Explain your reasoning.

d. Test the null hypothesis that there is no relationship between the variables against the alternative that there is an inverse relationship between them, at the $\alpha = 0.05$ level of significance. Discuss your conclusion. Be sure to include in your answer the test statistic used, the rejection region, and the *p*-value.

e. Test the hypothesis that for each percent higher for the relative price that the motel in question charges, it loses 1% of its occupancy rate. Formulate the null and alternative hypotheses in terms of the model parameters, carry out the relevant test at the 5% level of significance, and state your conclusion. Be sure to state the test statistic used, the rejection region, and the *p*-value.

3.29 We introduced Tennessee's Project STAR (Student/Teacher Achievement Ratio) in Exercise 2.22. The data file is *star5_small*. [The data file *star5* contains more observations and more variables.] Three types of classes were considered: small classes [$SMALL = 1$], regular-sized classes with a teacher aide [$AIDE = 1$], and regular-sized classes [$REGULAR = 1$].

a. Compute the sample mean and standard deviation for student math scores, *MATHSCORE*, in small classes. Compute the sample mean and standard deviation for student math scores, *MATHSCORE*, in regular classes, with no teacher aide. Which type of class had the higher average score? What is the difference in sample average scores for small classes versus regular-sized classes? Which type of class had the higher score standard deviation?

b. Consider students only in small classes or regular-sized classes without a teacher aide. Estimate the regression model $MATHSCORE = \beta_1 + \beta_2 SMALL + e$. How do the estimates of the regression parameters relate to the sample average scores calculated in part (a)?

c. Using the model from part (b), construct a 95% interval estimate of the expected *MATHSCORE* for a student in a regular-sized class and a student in a small class. Are the intervals fairly narrow or not? Do the intervals overlap?

d. Test the null hypothesis that the expected mathscore is no different in the two types of classes versus the alternative that expected *MATHSCORE* is higher for students in small classes using the 5% level of significance. State these hypotheses in terms of the model parameters, clearly state the test statistic you use, and the test rejection region. Calculate the *p*-value for the test. What is your conclusion?

e. Test the null hypothesis that the expected *MATHSCORE* is 15 points higher for students in small classes versus the alternative that it is not 15 points higher using the 10% level of significance. State these hypotheses in terms of the model parameters, clearly state the test statistic you use, and the test rejection region. Calculate the *p*-value for the test. What is your conclusion?

3.30 We introduced Tennessee's Project STAR (Student/Teacher Achievement Ratio) in Exercise 2.22. The data file is *star5_small*. [The data file *star5* contains more observations and more variables.] Three types of classes were considered: small classes [$SMALL = 1$], regular-sized classes with a teacher aide [$AIDE = 1$], and regular-sized classes [$REGULAR = 1$].

a. Compute the sample mean and standard deviation for student math scores, *MATHSCORE*, in regular classes with no teacher aide. Compute the sample mean and standard deviation for student math scores, *MATHSCORE*, in regular classes with a teacher aide. Which type of class had the higher average score? What is the difference in sample average scores for regular-sized classes versus regular sized classes with a teacher aide? Which type of class had the higher score standard deviation?

b. Consider students only in regular sized classes without a teacher aide and regular sized classes with a teacher aide. Estimate the regression model $MATHSCORE = \beta_1 + \beta_2 AIDE + e$. How do the estimates of the regression parameters relate to the sample average scores calculated in part (a)?

c. Using the model from part (b), construct a 95% interval estimate of the expected *MATHSCORE* for a student in a regular-sized class without a teacher aide and a regular-sized class with a teacher aide. Are the intervals fairly narrow or not? Do the intervals overlap?

d. Test the null hypothesis that the expected *MATHSCORE* is no different in the two types of classes versus the alternative that expected *MATHSCORE* is higher for students in regular-sized classes with a teacher aide, using the 5% level of significance. State these hypotheses in terms of the model parameters, clearly state the test statistic you use, and the test rejection region. Calculate the *p*-value for the test. What is your conclusion?

e. Test the null hypothesis that the expected *MATHSCORE* is three points, or more, higher for students in regular-sized classes with a teacher aide versus the alternative that the difference is less than three points, using the 10% level of significance. State these hypotheses in terms of the model parameters, clearly state the test statistic you use and the test rejection region. Calculate the *p*-value for the test. What is your conclusion?

3.31 Data on weekly sales of a major brand of canned tuna by a supermarket chain in a large midwestern U.S. city during a mid-1990s calendar year are contained in the data file *tuna*. There are 52 observations for each of the variables. The variable $SAL1$ = unit sales of brand no. 1 canned tuna, and $APR1$ = price per can of brand no. 1 tuna (in dollars).

a. Calculate the summary statistics for $SAL1$ and $APR1$. What are the sample means, minimum and maximum values, and their standard deviations. Plot each of these variables versus *WEEK*. How much variation in sales and price is there from week to week?

b. Plot the variable $SAL1$ (y-axis) against $APR1$ (x-axis). Is there a positive or inverse relationship? Is that what you expected, or not? Why?

c. Create the variable $PRICE1 = 100APR1$. Estimate the linear regression $SAL1 = \beta_1 + \beta_2 PRICE1 + e$. What is the point estimate for the effect of a one cent increase in the price of brand no. 1 on the sales of brand no. 1? What is a 95% interval estimate for the effect of a one cent increase in the price of brand no. 1 on the sales of brand no. 1?

d. Construct a 90% interval estimate for the expected number of cans sold in a week when the price per can is 70 cents.

e. Construct a 95% interval estimate of the elasticity of sales of brand no. 1 with respect to the price of brand no. 1 "at the means." Treat the sample means as constants and not random variables. Do you find the sales are fairly elastic, or fairly inelastic, with respect to price? Does this make economic sense? Why?

f. Test the hypothesis that elasticity of sales of brand no. 1 with respect to the price of brand no. 1 from part (e) is minus three against the alternative that the elasticity is not minus three. Use the 10% level of significance. Clearly, state the null and alternative hypotheses in terms of the model parameters, give the rejection region, and the *p*-value for the test. What is your conclusion?

3.32 What is the relationship between crime and punishment? We use data from 90 North Carolina counties to examine the question. County crime rates and other characteristics are observed over the period 1981–1987. The data are in the file *crime*. Use the 1985 data for this exercise.

a. Calculate the summary statistics for *CRMRTE* (crimes committed per person) and *PRBARR* (the probability of arrest = the ratio of arrests to offenses), including the maximums and minimums. Does there appear to be much variation from county to county in these variables?

b. Plot *CRMRTE* versus *PRBARR*. Do you observe a relationship between these variables?

c. Estimate the linear regression model $CRMRTE = \beta_1 + \beta_2 PRBARR + e$. If we increase the probability of arrest by 10% what will be the effect on the crime rate? What is a 95% interval estimate of this quantity?

d. Test the null hypothesis that there is no relationship between the county crime rate and the probability of arrest versus the alternative that there is an inverse relationship. State the null and alternative hypotheses in terms of the model parameters. Clearly, state the test statistic and its distribution if the null hypothesis is true and the test rejection region. Use the 1% level of significance. What is your conclusion?

Appendix 3A Derivation of the *t*-Distribution

Interval estimation and hypothesis testing procedures in this chapter involve the *t*-distribution. Here we develop the key result.

The first result that is needed is the normal distribution of the least squares estimator. Consider, for example, the normal distribution of b_2 the least squares estimator of β_2, which we denote as

$$b_2|\mathbf{x} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

A standardized normal random variable is obtained from b_2 by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2|\mathbf{x})}} \sim N(0,1) \tag{3A.1}$$

That is, the standardized random variable Z is normally distributed with mean 0 and variance 1. Despite the fact that the distribution of the least squares estimator b_2 depends on $\mathbf{x}$, the standardization leaves us with a pivotal statistic whose distribution depends on neither unknown parameters nor $\mathbf{x}$.

The second piece of the puzzle involves a chi-square random variable. If assumption SR6 holds, then the random error term e_i has a conditional normal distribution, $e_i|\mathbf{x} \sim N(0, \sigma^2)$. Standardize the random variable by dividing by its standard deviation so that $e_i/\sigma \sim N(0,1)$. The square of a standard normal random variable is a chi-square random variable (see Appendix B.5.2) with one degree of freedom, so $(e_i/\sigma)^2 \sim \chi^2_{(1)}$. If all the random errors are independent, then

$$\sum\left(\frac{e_i}{\sigma}\right)^2 = \left(\frac{e_1}{\sigma}\right)^2 + \left(\frac{e_2}{\sigma}\right)^2 + \cdots + \left(\frac{e_N}{\sigma}\right)^2 \sim \chi^2_{(N)} \tag{3A.2}$$

Since the true random errors are unobservable, we replace them by their sample counterparts, the least squares residuals $\hat{e}_i = y_i - b_1 - b_2 x_i$, to obtain

$$V = \frac{\sum \hat{e}_i^2}{\sigma^2} = \frac{(N-2)\hat{\sigma}^2}{\sigma^2} \tag{3A.3}$$

The random variable V in (3A.3) does not have a $\chi^2_{(N)}$ distribution because the least squares residuals are *not* independent random variables. All N residuals $\hat{e}_i = y_i - b_1 - b_2 x_i$ depend on the least squares estimators b_1 and b_2. It can be shown that only $N-2$ of the least squares residuals are independent in the simple linear regression model. Consequently, the random variable in (3A.3) has a chi-square distribution with $N-2$ degrees of freedom. That is, when multiplied by the constant $(N-2)/\sigma^2$, the random variable $\hat{\sigma}^2$ has a *chi-square distribution with* $N-2$ *degrees of freedom*,

$$V = \frac{(N-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-2)} \tag{3A.4}$$

The random variable V has a distribution that depends only on the degrees of freedom, $N-2$. Like Z in (3A.1), V is a pivotal statistic. We have *not* established the fact that the chi-square random variable V is statistically independent of the least squares estimators b_1 and b_2, but it is. The proof is beyond the scope of this book. Consequently, V and the standard normal random variable Z in (3A.1) are independent.

From the two random variables V and Z, we can form a t-random variable. A t-random variable is formed by dividing a standard normal random variable, $Z \sim N(0,1)$, by the square root of an *independent* chi-square random variable, $V \sim \chi^2_{(m)}$, that has been divided by its degrees of freedom, m. That is,

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)} \tag{3A.5}$$

The t-distribution's shape is completely determined by the degrees of freedom parameter, m, and the distribution is symbolized by $t_{(m)}$. See Appendix B.5.3. Using Z and V from (3A.1) and (3A.4), respectively, we have

$$t = \frac{Z}{\sqrt{V/(N-2)}} = \frac{(b_2 - \beta_2)\Big/\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}}{\sqrt{\dfrac{(N-2)\hat{\sigma}^2/\sigma^2}{N-2}}}$$

$$= \frac{b_2 - \beta_2}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}} = \frac{b_2 - \beta_2}{\sqrt{\widehat{\mathrm{var}}(b_2)}} = \frac{b_2 - \beta_2}{\mathrm{se}(b_2)} \sim t_{(N-2)} \tag{3A.6}$$

The second line is the key result that we state in (3.2), with its generalization in (3.3).

Appendix 3B Distribution of the t-Statistic under H_1

To better understand how t-tests work, let us examine the t-statistic in (3.7) when the null hypothesis is not true. We can do that by writing it out in some additional detail. What happens to Z in (3A.1) if we test a hypothesis $H_0: \beta_2 = c$ that might not be true? Instead of subtracting β_2, we subtract c, to obtain

$$\frac{b_2 - c}{\sqrt{\mathrm{var}(b_2)}} = \frac{b_2 - \beta_2 + \beta_2 - c}{\sqrt{\mathrm{var}(b_2)}} = \frac{b_2 - \beta_2}{\sqrt{\mathrm{var}(b_2)}} + \frac{\beta_2 - c}{\sqrt{\mathrm{var}(b_2)}} = Z + \delta \sim N(\delta, 1)$$

The statistic we obtain is the standard normal Z plus another factor, $\delta = (\beta_2 - c)\Big/\sqrt{\text{var}(b_2)}$, that is zero only if the null hypothesis is true. A **noncentral *t*-random variable** is formed from the ratio

$$t|\mathbf{x} = \frac{Z + \delta}{\sqrt{V/m}} \sim t_{(m,\delta)} \tag{3B.1}$$

This is a more general t-statistic, with m degrees of freedom and noncentrality parameter δ, denoted $t_{(m,\,\delta)}$. It has a distribution that is not centered at zero unless $\delta = 0$. The non-central t-distribution is introduced in Appendix B.7.3. It is the factor δ that leads the t-test to reject a false null hypothesis with probability greater than α, which is the probability of a Type I error. Because δ depends on the sample data, we have indicated that the non-central t-distribution is conditional on $\mathbf{x}$. If the null hypothesis is true then $\delta = 0$ and the t-statistic does not depend on any unknown parameters or $\mathbf{x}$; it is a pivotal statistic.

Suppose that we have a sample of size $N = 40$ so that the degrees of freedom are $N - 2 = 38$ and we test a hypothesis concerning β_2 such that $\beta_2 - c = 1$. Using a right-tail test, the probability of rejecting the null hypothesis is $P(t > 1.686)$, where $t_{(0.95,\,38)} = 1.686$ is from Statistical Table 2, the percentiles of the usual t-distribution. If $\delta = 0$, this rejection probability is 0.05. With $\beta_2 - c = 1$, we must compute the right-tail probability using the non-central t-distribution with noncentrality parameter

$$\delta = \frac{\beta_2 - c}{\sqrt{\text{var}(b_2)}} = \frac{\beta_2 - c}{\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}} = \frac{\sqrt{\sum(x_i - \bar{x})^2}(\beta_2 - c)}{\sigma} \tag{3B.2}$$

For a numerical example, we use values arising from the simulation experiment used in Appendix 2H. The sample of x-values consists of $x_i = 10, i = 1, \ldots, 20$ and $x_i = 20$, $i = 21, \ldots, 40$. The sample mean is $\bar{x} = 15$ so that $\sum(x_i - \bar{x})^2 = 40 \times 5^2 = 1000$. Also, $\sigma^2 = 2500$. The noncentrality parameter is

$$\delta = \frac{\sqrt{\sum(x_i - \bar{x})^2}(\beta_2 - c)}{\sigma} = \frac{\sqrt{1000}(\beta_2 - c)}{\sqrt{2500}} = 0.63246(\beta_2 - c)$$

Thus, the probability of rejecting the null hypothesis $H_0 : \beta_2 = 9$ versus $H_1 : \beta_2 > 9$ when the true value of $\beta_2 = 10$ is

$$P\left(t_{(38,\,0.63246)} > 1.686\right) = 1 - P\left(t_{(38,\,0.63246)} \leq 1.686\right) = 0.15301$$

The probability calculation uses the cumulative distribution function for the non-central t-distribution, which is available in econometric software and at some websites. Similarly, the probability of rejecting the null hypothesis $H_0 : \beta_2 = 8$ versus $H_1 : \beta_2 > 8$ when the true value of $\beta_2 = 10$ is

$$P\left(t_{(38,\,1.26491)} > 1.686\right) = 1 - P\left(t_{(38,\,1.26491)} \leq 1.686\right) = 0.34367$$

Why does the probability of rejection increase? The effect of the noncentrality parameter is to shift the t-distribution rightward, as shown in Appendix B.7.3. For example, the probability of rejecting the null hypothesis $H_0 : \beta_2 = 9$ versus $H_1 : \beta_2 > 9$ is shown in Figure 3B.1.

The solid curve is the usual central t-distribution with 38 degrees of freedom. The area under the curve to the right of 1.686 is 0.05. The dashed curve is the non-central t-distribution with $\delta = 0.63246$. The area under the curve to the right of 1.686 is larger, approximately 0.153.

The probability of rejecting a false null hypothesis is called a test's **power**. In an ideal world, we would reject false null hypotheses always, and if we had an infinite amount of data we could. The keys to a t-test's power are the three ingredients making the noncentrality parameter larger.

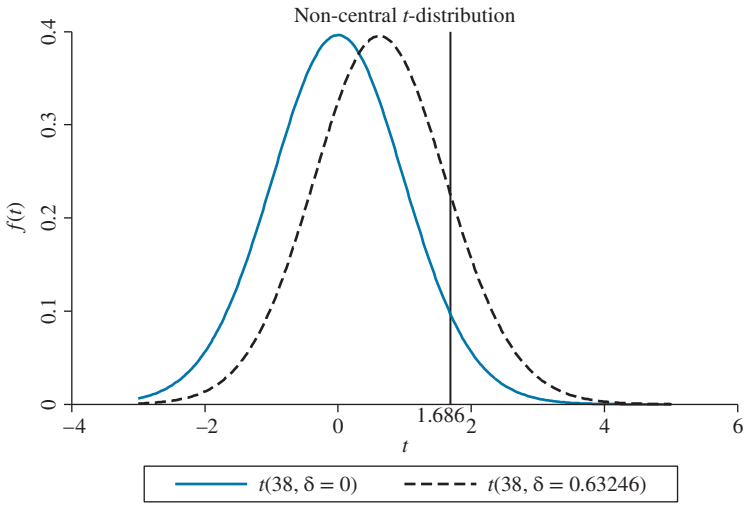


FIGURE 3B.1 **Probability of rejecting $H_0: \beta_2 = 9$.**

A larger noncentrality parameter shifts the t-distribution further rightward and increases the probability of rejection. Thus, the probability of rejecting a false null hypothesis increases when

1. The magnitude of the hypothesis error $\beta_2 - c$ increases.
2. The smaller the true error variance, σ^2, that measures the overall model uncertainty.
3. The larger the total variation in the explanatory variable, which might be the result of a larger sample size.

In a real situation, the actual power of a test is unknown because we do not know β_2 or σ^2, and the power calculation depends on being given the x-values. Nevertheless, it is good to know the factors that will increase the probability of rejecting a false null hypothesis. In the following section, we carry out a Monte Carlo simulation experiment to illustrate the power calculations above.

Recall that a Type II error is failing to reject a hypothesis that is false. Consequently, the probability of a Type II error is the complement of the test's power. For example, the probability of a Type II error when testing $H_0: \beta_2 = 9$ versus $H_1: \beta_2 > 9$ when the true value of $\beta_2 = 10$ is

$$P\left(t_{(38,\,0.63246)} \leq 1.686\right) = 1 - 0.15301 = 0.84699$$

For testing $H_0: \beta_2 = 8$ versus $H_1: \beta_2 > 8$, when the true value is $\beta_2 = 10$, the probability of a Type II error is $P\left(t_{(38,\,1.26491)} \leq 1.686\right) = 1 - 0.34367 = 0.65633$. As test power increases, the probability of a Type II error falls, and vice versa.

Appendix 3C Monte Carlo Simulation

In Appendix 2H, we introduced a Monte Carlo simulation to illustrate the repeated sampling properties of the least squares estimators. In this appendix, we use the same framework to illustrate the repeated sampling performances of interval estimators and hypothesis tests.

Recall that the **data generation process** for the simple linear regression model is given by

$$y_i = E\left(y_i|x_i\right) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

The Monte Carlo parameter values are $\beta_1 = 100$ and $\beta_2 = 10$. The value of x_i is 10 for the first 20 observations and 20 for the remaining 20 observations, so that the regression functions are

$$E(y_i|x_i = 10) = 100 + 10x_i = 100 + 10 \times 10 = 200, \quad i = 1, \ldots, 20$$

$$E(y_i|x_i = 20) = 100 + 10x_i = 100 + 10 \times 20 = 300, \quad i = 21, \ldots, 40$$

The random errors are independently and normally distributed with mean 0 and variance $\text{var}(e_i|x_i) = \sigma^2 = 2{,}500$, or $e_i|x \sim N(0, 2500)$.

When studying the performance of hypothesis tests and interval estimators, it is necessary to use enough Monte Carlo samples so that the percentages involved are estimated precisely enough to be useful. For tests with probability of Type I error $\alpha = 0.05$, we should observe true null hypotheses being rejected 5% of the time. For 95% interval estimators, we should observe that 95% of the interval estimates contain the true parameter values. We use $M = 10{,}000$ Monte Carlo samples so that the experimental error is very small. See Appendix 3C.3 for an explanation.

3C.1 Sampling Properties of Interval Estimators

In Appendix 2H.4, we created one sample of data that is in the data file *mcl_fixed_x*. The least squares estimates using these data values are

$$\hat{y} = 127.2055 + 8.7325x$$
$$(23.3262) \quad (1.4753)$$

A 95% interval estimate of the slope is $b_2 \pm t_{(0.975,\,38)}\text{se}(b_2) = [5.7460, 11.7191]$. We see that for this sample, the 95% interval estimate contains the true slope parameter value $\beta_2 = 10$.

We repeat the process of estimation and interval estimation 10,000 times. In these repeated samples, 95.03% of the interval estimates contain the true parameter. Table 3C.1 contains results for the Monte Carlo samples 321–330 for illustration purposes. The estimates are *B2*, the standard error is *SE*, the lower bound of the 95% interval estimate is *LB*, and the upper bound is *UB*. The variable *COVER* = 1 if the interval estimate contains the true parameter value. Two of the intervals do not contain the true parameter value $\beta_2 = 10$. The 10 sample results we are reporting were chosen to illustrate that interval estimates do not cover the true parameter in all cases.

The lesson is, that in many samples from the data generation process, and if assumptions SR1–SR6 hold, the procedure for constructing 95% interval estimates "works" 95% of the time.

TABLE 3C.1 Results of 10000 Monte Carlo Simulations

SAMPLE	B2	SE	TSTAT	REJECT	LB	UB	COVER
321	7.9600	1.8263	−1.1170	0	4.2628	11.6573	1
322	11.3093	1.6709	0.7836	0	7.9267	14.6918	1
323	9.8364	1.4167	−0.1155	0	6.9683	12.7044	1
324	11.4692	1.3909	1.0563	0	8.6535	14.2849	1
325	9.3579	1.5127	−0.4245	0	6.2956	12.4202	1
326	9.6332	1.5574	−0.2355	0	6.4804	12.7861	1
327	9.0747	1.2934	−0.7154	0	6.4563	11.6932	1
328	7.0373	1.3220	−2.2411	0	4.3611	9.7136	0
329	13.1959	1.7545	1.8215	1	9.6441	16.7478	1
330	14.4851	2.1312	2.1046	1	10.1708	18.7994	0

3C.2 Sampling Properties of Hypothesis Tests

The null hypothesis $H_0: \beta_2 = 10$ is true. If we use the one-tail alternative $H_1: \beta_2 > 10$ and level of significance $\alpha = 0.05$, the null hypothesis is rejected if the test statistic $t = (b_2 - 10)/\text{se}(b_2) > 1.68595$, which is the 95th percentile of the t-distribution with 38 degrees of freedom.[3] For the sample *mc1_fixed_x*, the calculated value of the t-statistic is -0.86, so we fail to reject the null hypothesis, which in this case is the correct decision.

We repeat the process of estimation and hypothesis testing 10,000 times. In these samples, 4.98% of the tests reject the null hypothesis that the parameter value is 10. In Table 3C.1, the t-statistic value is *TSTAT* and *REJECT* = 1 if the null hypothesis is rejected. We see that samples 329 and 330 incorrectly reject the null hypothesis.

The lesson is that in many samples from the data generation process, and if assumptions SR1–SR6 hold, the procedure for testing a true null hypothesis at significance level $\alpha = 0.05$ rejects the true null hypothesis 5% of the time. Or, stated positively, the test procedure does not reject the true null hypothesis 95% of the time.

To investigate the power of the t-test, the probability that it rejects a false hypothesis, we tested $H_0: \beta_2 = 9$ versus $H_1: \beta_2 > 9$ and $H_0: \beta_2 = 8$ versus $H_1: \beta_2 > 8$. The theoretical rejection rates we calculated in Appendix 3B are 0.15301 in the first case and 0.34367 in the second. In 10,000 Monte Carlo samples, the first hypothesis was rejected in 1515 samples for a rejection rate of 0.1515. The second hypothesis was rejected in 3500 of the samples, a rejection rate of 0.35. The Monte Carlo values are very close to the true rejection rates.

3C.3 Choosing the Number of Monte Carlo Samples

A 95% confidence interval estimator should contain the true parameter value 95% of the time in many samples. The M samples in a Monte Carlo experiment are independent experimental trials in which the probability of a "success," an interval containing the true parameter value, is $P = 0.95$. The number of successes follows a **binomial** distribution. The **proportion** of successes $\hat{P}$ in M trials is a random variable with expectation P and variance $P(1-P)/M$. If the number of Monte Carlo samples M is large, a 95% interval estimate of the proportion of Monte Carlo successes is $P \pm 1.96\sqrt{P(1-P)/M}$. If $M = 10{,}000$, this interval is [0.9457, 0.9543]. We chose $M = 10{,}000$ so that this interval would be narrow, giving us confidence that *if* the true probability of success is 0.95 we will obtain a Monte Carlo average close to 0.95 with a "high" degree of confidence. Our result that 95.03% of our interval estimates contain the true parameter β_2 is "within" the margin of error for such Monte Carlo experiments. On the other hand, if we had used $M = 1000$ Monte Carlo samples, the interval estimate of the proportion of Monte Carlo successes would be, [0.9365, 0.9635]. With this wider interval, the proportion of Monte Carlo successes could be quite different from 0.95, casting a shadow of doubt on whether our method was working as advertised or not.

Similarly, for a test with probability of rejection $\alpha = 0.05$, the 95% interval estimate of the proportion of Monte Carlo samples leading to rejection is $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/M}$. If $M = 10{,}000$, this interval is [0.0457, 0.0543]. That our Monte Carlo experiments rejected the null hypothesis 4.98% of the time is within this margin of error. If we had chosen $M = 1000$, then the proportion of Monte Carlo rejections is estimated to be in the interval [0.0365, 0.0635], which again leaves just a little too much wiggle room for comfort.

The point is that if fewer Monte Carlo samples are chosen the "noise" in the Monte Carlo experiment can lead to a percent of successes or rejections that has too wide a margin of error for

[3] We use a t-critical value with more decimals, instead of the table value 1.686, to ensure accuracy in the Monte Carlo experiment.

us to tell whether the statistical procedure, interval estimation, or hypothesis testing is "working" properly or not.[4]

3C.4 Random-*x* Monte Carlo Results

We used the "fixed-*x*" framework in Monte Carlo results reported in Sections 3C.1 and 3C.2. In each Monte Carlo sample, the *x*-values were $x_i = 10$ for the first 20 observations and $x_i = 20$ for the next 20 observations. Now we modify the experiment to the random-*x* case, as in Appendix 2H.7. The data-generating equation remains $y_i = 100 + 10x_i + e_i$ with the random errors having a normal distribution with mean zero and standard deviation 50, $e_i \sim N(0, 50^2 = 2500)$. We randomly choose *x*-values from a normal distribution with mean $\mu_x = 15$ and standard deviation $\sigma_x = 1.6$, so $x \sim N(15, 1.6^2 = 2.56)$.

One sample of data is in the data file *mc1_random_x*. Using these values, we obtain the least squares estimates

$$\hat{y} = 116.7410 + 9.7628x$$
$$(84.7107)\ (5.5248)$$

A 95% interval estimate of the slope is $b_2 \pm t_{(0.975,\,38)}\text{se}(b_2) = [-1.4216, 20.9472]$. For this sample, the 95% interval estimate contains the true slope parameter value $\beta_2 = 10$.

We generate 10,000 Monte Carlo samples using this design and compute the least squares estimates and 95% interval estimates. In these samples, with **x** varying from sample to sample, the 95% interval estimates for β_2 contain the true value in 94.87% of the samples. Table 3C.2 contains results for the Monte Carlo samples 321–330 for illustration purposes. The estimates are *B2*, the standard error *SE*, the lower bound of the 95% interval estimate is *LB*, and the upper bound is *UB*. The variable *COVER* = 1 if the interval contains the true parameter value. In the selected samples, one interval estimate, 323, does not contain the true parameter value.

In the Monte Carlo experiment, we test the null hypothesis $H_0: \beta_2 = 10$ against the alternative $H_1: \beta_2 > 10$ using the *t*-statistic $t = (b_2 - 10)/\text{se}(b_2)$. We reject the null hypothesis if $t \geq 1.685954$, which is the 95th percentile of the $t_{(38)}$ distribution. In Table 3C.2, the *t*-statistic values are *TSTAT* and *REJECT* = 1 if the test rejects the null hypothesis. In 5.36% of the 10,000 Monte Carlo samples, we reject the null hypothesis, which is within the margin of error discussed in Section 3C.2. In Table 3C.2, for sample 323, the true null hypothesis was rejected.

TABLE 3C.2 Results of 10,000 Monte Carlo Simulations with Random-*x*

SAMPLE	*B2*	*SE*	*TSTAT*	*REJECT*	*LB*	*UB*	*COVER*
321	9.6500	5.1341	−0.0682	0	−0.7434	20.0434	1
322	7.4651	4.3912	−0.5773	0	−1.4244	16.3547	1
323	22.9198	5.6616	2.2820	1	11.4584	34.3811	0
324	8.6675	4.8234	−0.2763	0	−1.0970	18.4320	1
325	18.7736	5.2936	1.6574	0	8.0573	29.4899	1
326	16.4197	3.8797	1.6547	0	8.5657	24.2738	1
327	3.7841	5.1541	−1.2060	0	−6.6500	14.2181	1
328	3.6013	4.9619	−1.2896	0	−6.4436	13.6462	1
329	10.5061	5.6849	0.0890	0	−1.0024	22.0145	1
330	9.6342	4.8478	−0.0755	0	−0.1796	19.4481	1

[4]Other details concerning Monte Carlo simulations can be found in *Microeconometrics: Methods and Applications*, by A. Colin Cameron and Pravin K. Trivedi (Cambridge University Press, 2005). The material is advanced.

We conclude from these simulations that in the random-x cases there is no evidence that inferences do not perform as expected, with 95% of intervals covering the true parameter value and 5% of tests rejecting a true null hypothesis.

To investigate the power of the t-test, the probability that it rejects a false hypothesis, we tested $H_0: \beta_2 = 9$ versus $H_1: \beta_2 > 9$ and $H_0: \beta_2 = 8$ versus $H_1: \beta_2 > 8$. In 10,000 Monte Carlo samples, the first hypothesis was rejected in 7.8% of the time and the second hypothesis was rejected 11.15% of the time. These rejection rates are far less than in the fixed-x results studied in Appendix 3B and less than the empirical rejection rates in the simulation results in Appendix 3C.2. We noted that the ability of the t-test to reject a false hypothesis was related to the magnitude of the noncentrality parameter in (3A.8), $\delta = \sqrt{\sum(x_i - \bar{x})^2}(\beta_2 - c)/\sigma$. In these experiments, the factors $(\beta_2 - c) = 1$ and 2 and $\sigma = 50$ are the same as in the fixed-x example. What must have changed? The only remaining factor is the variation in the x-values, $\sum(x_i - \bar{x})^2$. In the earlier example, $\sum(x_i - \bar{x})^2 = 1000$ and the x-values were fixed in repeated samples. In this experiment, the x-values were not fixed but random, and for each sample of x-values, the amount of variation changes. We specified the variance of x to be 2.56, and in 10,000 Monte Carlo experiments, the average of the sample variance $s_x^2 = 2.544254$ and the average of the variation in x about its mean, $\sum(x_i - \bar{x})^2$, was 99.22591, or about one-tenth the variation in the fixed-x case. It is perfectly clear why the power of the test in the random-x case was lower, it is because on average $\sum(x_i - \bar{x})^2$ was smaller.

CHAPTER 4

Prediction, Goodness-of-Fit, and Modeling Issues

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain how to use the simple linear regression model to predict the value of y for a given value of x.
2. Explain, intuitively and technically, why predictions for x values further from $\bar{x}$ are less reliable.
3. Explain the meaning of *SST*, *SSR*, and *SSE*, and how they are related to R^2.
4. Define and explain the meaning of the coefficient of determination.
5. Explain the relationship between correlation analysis and R^2.
6. Report the results of a fitted regression equation in such a way that confidence intervals and hypothesis tests for the unknown coefficients can be constructed quickly and easily.
7. Describe how estimated coefficients and other quantities from a regression equation will change when the variables are scaled. Why would you want to scale the variables?
8. Appreciate the wide range of nonlinear functions that can be estimated using a model that is linear in the parameters.
9. Write down the equations for the log-log, log-linear, and linear-log functional forms.
10. Explain the difference between the slope of a functional form and the elasticity from a functional form.
11. Explain how you would go about choosing a functional form and deciding that a functional form is adequate.
12. Explain how to test whether the equation ''errors'' are normally distributed.
13. Explain how to compute a prediction, a prediction interval, and a goodness-of-fit measure in a log-linear model.
14. Explain alternative methods for detecting unusual, extreme, or incorrect data values.

KEYWORDS

coefficient of determination
correlation
forecast error
functional form
goodness-of-fit
growth model
influential observations
Jarque–Bera test
kurtosis
least squares predictor
linear model
linear relationship
linear-log model
log-linear model
log-log model
log-normal distribution
prediction
prediction interval
R^2
residual diagnostics
scaling data
skewness
standard error of the forecast

In Chapter 3, we focused on making statistical inferences, constructing confidence intervals, and testing hypotheses about regression parameters. Another purpose of the regression model, and the one we focus on first in this chapter, is **prediction**. A prediction is a forecast of an unknown value of the dependent variable y given a particular value of x. A **prediction interval**, much like a confidence interval, is a range of values in which the unknown value of y is likely to be located. Examining the **correlation** between sample values of y and their predicted values provides a **goodness-of-fit** measure called $\boldsymbol{R^2}$ that describes how well our model fits the data. For each observation in the sample, the difference between the predicted value of y and the actual value is a **residual**. Diagnostic measures constructed from the residuals allow us to check the adequacy of the **functional form** used in the regression analysis and give us some indication of the validity of the regression assumptions. We will examine each of these ideas and concepts in turn.

4.1 Least Squares Prediction

In Example 2.4, we briefly introduced the idea that the least squares estimates of the linear regression model provide a way to predict the value of y for any value of x. The ability to predict is important to business economists and financial analysts who attempt to forecast the sales and revenues of specific firms; it is important to government policymakers who attempt to predict the rates of growth in national income, inflation, investment, saving, social insurance program expenditures, and tax revenues; and it is important to local businesses who need to have predictions of growth in neighborhood populations and income so that they may expand or contract their provision of services. Accurate predictions provide a basis for better decision making in every type of planning context. In this section, we explore the use of linear regression as a tool for prediction.

Given the simple linear regression model and assumptions SR1–SR6, let x_0 be a given value of the explanatory variable. We want to predict the corresponding value of y, which we call y_0. In order to use regression analysis as a basis for prediction, we must assume that y_0 and x_0 are related to one another by the same regression model that describes our sample of data, so that, in particular, SR1 holds for these observations

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \tag{4.1}$$

where e_0 is a random error. We assume that $E(y_0|x_0) = \beta_1 + \beta_2 x_0$ and $E(e_0) = 0$. We also assume that e_0 has the same variance as the regression errors, $\text{var}(e_0) = \sigma^2$, and e_0 is uncorrelated with the random errors that are part of the sample data, so that $\text{cov}(e_0, e_i|\mathbf{x}) = 0$, $i = 1, 2, \ldots, N$.

The task of **predicting** y_0 is related to the problem of **estimating** $E(y_0|x_0) = \beta_1 + \beta_2 x_0$, which we discussed in Section 3.6. The outcome $y_0 = E(y_0|x_0) + e_0 = \beta_1 + \beta_2 x_0 + e_0$ is composed of two parts, the systematic, nonrandom part $E(y_0|x_0) = \beta_1 + \beta_2 x_0$, and a random

component e_0. We estimate the systematic portion using $\hat{E}(y_0|x_0) = b_1 + b_2x_0$ and add an "estimate" of e_0 equal to its expected value, which is zero. Therefore, the prediction $\hat{y}_0$ is given by $\hat{y}_0 = \hat{E}(y_0|x_0) + 0 = b_1 + b_2x_0$. Despite the fact that we use the same statistic for both $\hat{y}_0$ and $\hat{E}(y_0|x_0)$, we distinguish between them because, although $E(y_0|x_0) = \beta_1 + \beta_2x_0$ is not random, the outcome y_0 is random. Consequently, as we will see, there is a difference between the **interval estimate** of $E(y_0|x_0) = \beta_1 + \beta_2x_0$ and the **prediction interval** for y_0.

Following from the discussion in the previous paragraph, the **least squares predictor** of y_0 comes from the fitted regression line

$$\hat{y}_0 = b_1 + b_2x_0 \tag{4.2}$$

That is, the predicted value $\hat{y}_0$ is given by the point on the least squares fitted line where $x = x_0$, as shown in Figure 4.1. How good is this prediction procedure? The least squares estimators b_1 and b_2 are random variables—their values vary from one sample to another. It follows that the least squares predictor $\hat{y}_0 = b_1 + b_2x_0$ must also be random. To evaluate how well this predictor performs, we define the **forecast error**, which is analogous to the least squares residual,

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2x_0 + e_0) - (b_1 + b_2x_0) \tag{4.3}$$

We would like the forecast error to be small, implying that our forecast is close to the value we are predicting. Taking the conditional expected value of f, we find

$$\begin{aligned} E(f|\mathbf{x}) &= \beta_1 + \beta_2x_0 + E(e_0) - \left[E(b_1|\mathbf{x}) + E(b_2|\mathbf{x})\, x_0\right] \\ &= \beta_1 + \beta_2x_0 + 0 - \left[\beta_1 + \beta_2x_0\right] \\ &= 0 \end{aligned}$$

which means, on average, the forecast error is zero, and $\hat{y}_0$ is an **unbiased predictor** of y_0. However, unbiasedness does not necessarily imply that a particular forecast will be close to the actual value. The probability of a small forecast error also depends on the variance of the forecast error. Although we will not prove it, $\hat{y}_0$ is the **best linear unbiased predictor (BLUP)** of y_0 if assumptions SR1–SR5 hold. This result is reasonable given that the least squares estimators b_1 and b_2 are best linear unbiased estimators.

Using (4.3) and what we know about the variances and covariances of the least squares estimators, we can show (see Appendix 4A) that the variance of the forecast error is

$$\text{var}(f|\mathbf{x}) = \sigma^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{4.4}$$

Notice that some of the elements of this expression appear in the formulas for the variances of the least squares estimators and affect the precision of prediction in the same way that they affect the precision of estimation. We would prefer that the variance of the forecast error be small, which

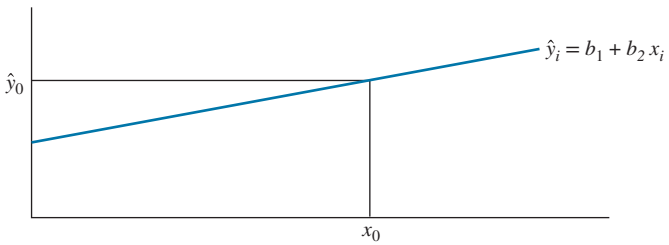


FIGURE 4.1 **A point prediction.**

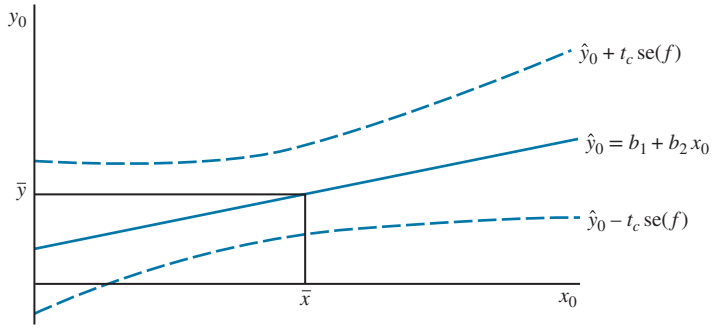


FIGURE 4.2 **Point and interval prediction.**

would increase the probability that the prediction $\hat{y}_0$ is close to the value y_0, we are trying to predict. Note that the variance of the forecast error is smaller when

- **i.** the overall uncertainty in the model is smaller, as measured by the variance of the random errors σ^2
- **ii.** the sample size N is larger
- **iii.** the variation in the explanatory variable is larger
- **iv.** the value of $(x_0 - \bar{x})^2$ is small

The new addition is the term $(x_0 - \bar{x})^2$, which measures how far x_0 is from the center of the x-values. The more distant x_0 is from the center of the sample data the larger the forecast variance will become. Intuitively, this means that we are able to do a better job predicting in the region where we have more sample information, and we will have less accurate predictions when we try to predict outside the limits of our data.

In practice we replace σ^2 in (4.4) by its estimator $\hat{\sigma}^2$ to obtain

$$\widehat{\mathrm{var}}(f|\mathbf{x}) = \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

The square root of this estimated variance is the **standard error of the forecast**

$$\mathrm{se}(f) = \sqrt{\widehat{\mathrm{var}}(f|\mathbf{x})} \tag{4.5}$$

Defining the critical value t_c to be the $100(1-\alpha/2)$-percentile from the t-distribution, we can obtain a $100(1-\alpha)\%$ **prediction interval** as

$$\hat{y}_0 \pm t_c \mathrm{se}(f) \tag{4.6}$$

See Appendix 4A for some details related to the development of this result.

Following our discussion of $\mathrm{var}(f|\mathbf{x})$ in (4.4), the farther x_0 is from the sample mean $\bar{x}$, the larger the variance of the prediction error will be, and the less reliable the prediction is likely to be. In other words, our predictions for values of x_0 close to the sample mean $\bar{x}$ are more reliable than our predictions for values of x_0 far from the sample mean $\bar{x}$. This fact shows up in the size of our prediction intervals. The relationship between point and interval predictions for different values of x_0 is illustrated in Figure 4.2. A point prediction is given by the fitted least squares line $\hat{y}_0 = b_1 + b_2 x_0$. The prediction interval takes the form of two bands around the fitted least squares line. Because the forecast variance increases the farther x_0 is from the sample mean $\bar{x}$, the confidence bands are their narrowest when $x_0 = \bar{x}$, and they increase in width as $|x_0 - \bar{x}|$ increases.

EXAMPLE 4.1 | Prediction in the Food Expenditure Model

In Example 2.4, we predicted that a household with $x_0 = \$2{,}000$ weekly income would spend \$287.61 on food using the calculation

$$\hat{y}_0 = b_1 + b_2 x_0 = 83.4160 + 10.2096(20) = 287.6089$$

Now we are able to attach a "confidence interval" to this prediction. The estimated variance of the forecast error is

$$\begin{aligned}\widehat{\text{var}}(f|\mathbf{x}) &= \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \\ &= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \\ &= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2\widehat{\text{var}}(b_2|\mathbf{x})\end{aligned}$$

In the last line, we have recognized the estimated variance of b_2 from (2.21). In Example 2.5 we obtained the values $\hat{\sigma}^2 = 8013.2941$ and $\widehat{\text{var}}(b_2|\mathbf{x}) = 4.3818$. For the food expenditure data, $N = 40$ and the sample mean of the explanatory variable is $\bar{x} = 19.6048$. Using these values, we obtain the standard error of the forecast $\text{se}(f) = \sqrt{\widehat{\text{var}}(f|\mathbf{x})} = \sqrt{8214.31} = 90.6328$. If we select $1 - \alpha = 0.95$, then $t_c = t_{(0.975,\,38)} = 2.0244$ and the 95% prediction interval for y_0 is

$$\begin{aligned}\hat{y}_0 \pm t_c\text{se}(f) &= 287.6069 \pm 2.0244(90.6328) \\ &= [104.1323,\ 471.0854]\end{aligned}$$

Our prediction interval suggests that a household with \$2,000 weekly income will spend somewhere between \$104.13 and \$471.09 on food. Such a wide interval means that our point prediction \$287.61 is not very reliable. We have obtained this wide prediction interval for the value of $x_0 = 20$ that is close to the sample mean $\bar{x} = 19.60$. For values of x that are more extreme, the prediction interval would be even wider. The unreliable predictions may be slightly improved if we collect a larger sample of data, which will improve the precision with which we estimate the model parameters. However, in this example the magnitude of the estimated error variance $\hat{\sigma}^2$ is very close to the estimated variance of the forecast error $\widehat{\text{var}}(f|\mathbf{x})$, indicating that the primary uncertainty in the forecast comes from large uncertainty in the model. This should not be a surprise, since we are predicting household behavior, which is a complicated phenomenon, on the basis of a single household characteristic, income. Although income is a key factor in explaining food expenditure, we can imagine that many other household demographic characteristics may play a role. To more accurately predict food expenditure, we may need to include these additional factors into the regression model. Extending the simple regression model to include other factors will begin in Chapter 5.

4.2 Measuring Goodness-of-Fit

Two major reasons for analyzing the model

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{4.7}$$

are to explain how the dependent variable (y_i) changes as the independent variable (x_i) changes and to predict y_0 given an x_0. These two objectives come under the broad headings of estimation and prediction. Closely allied with the prediction problem discussed in the previous section is the desire to use x_i to explain as much of the variation in the dependent variable y_i as possible. In the regression model (4.7), we call x_i the "explanatory" variable because we hope that its variation will "explain" the variation in y_i.

To develop a measure of the variation in y_i that is explained by the model, we begin by separating y_i into its explainable and unexplainable components. We have assumed that

$$y_i = E(y_i|\mathbf{x}) + e_i \tag{4.8}$$

where $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ is the explainable, "systematic" component of y_i, and e_i is the random, unsystematic, and unexplainable component of y_i. While both of these parts are unobservable to us, we can estimate the unknown parameters β_1 and β_2 and, analogous to (4.8), decompose the value of y_i into

$$y_i = \hat{y}_i + \hat{e}_i \tag{4.9}$$

where $\hat{y}_i = b_1 + b_2 x_i$ and $\hat{e}_i = y_i - \hat{y}_i$.

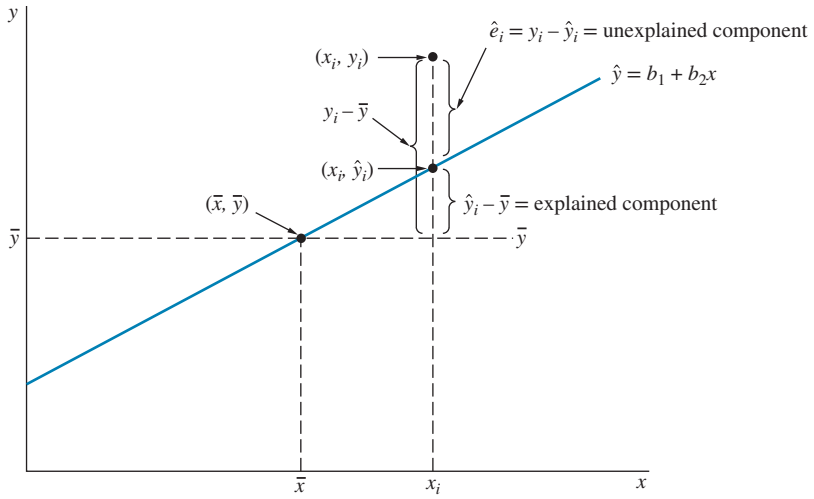


FIGURE 4.3 **Explained and unexplained components of y_i.**

In Figure 4.3, the "point of the means" $(\bar{x}, \bar{y})$ is shown, with the least squares fitted line passing through it. This is a characteristic of the least squares fitted line whenever the regression model includes an intercept term. Subtract the sample mean $\bar{y}$ from both sides of the equation to obtain

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i \tag{4.10}$$

As shown in Figure 4.3, the difference between y_i and its mean value $\bar{y}$ consists of a part that is "explained" by the regression model $\hat{y}_i - \bar{y}$ and a part that is unexplained $\hat{e}_i$.

The breakdown in (4.10) leads to a decomposition of the total sample variability in y into explained and unexplained parts. Recall from your statistics courses (see Appendix C4) that if we have a sample of observations $y_1, y_2, \ldots, y_N$, two descriptive measures are the sample mean $\bar{y}$ and the sample variance

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{N - 1}$$

The numerator of this quantity, the sum of squared differences between the sample values y_i and the sample mean $\bar{y}$, is a measure of the total variation in the sample values. If we square and sum both sides of (4.10) and use the fact that the cross-product term $\sum (\hat{y}_i - \bar{y})\hat{e}_i = 0$ (see Appendix 4B), we obtain

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2 \tag{4.11}$$

Equation (4.11) gives us a decomposition of the "total sample variation" in y into explained and unexplained components. Specifically, these "sums of squares" are as follows:

1. $\sum (y_i - \bar{y})^2$ = total sum of squares = *SST*: a measure of *total variation* in y about the sample mean.
2. $\sum (\hat{y}_i - \bar{y})^2$ = sum of squares due to the regression = *SSR*: that part of total variation in y, about the sample mean, that is explained by, or due to, the regression. Also known as the "explained sum of squares."
3. $\sum \hat{e}_i^2$ = sum of squares due to error = *SSE*: that part of total variation in y about its mean that is not explained by the regression. Also known as the unexplained sum of squares, the residual sum of squares, or the sum of squared errors.

Using these abbreviations, (4.11) becomes

$$SST = SSR + SSE$$

This decomposition of the total variation in y into a part that is explained by the regression model and a part that is unexplained allows us to define a measure, called the **coefficient of determination**, or R^2, that is the proportion of variation in y explained by x within the regression model.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{4.12}$$

The closer R^2 is to 1, the closer the sample values y_i are to the fitted regression equation $\hat{y}_i = b_1 + b_2 x_i$. If $R^2 = 1$, then all the sample data fall exactly on the fitted least squares line, so $SSE = 0$, and the model fits the data "perfectly." If the sample data for y and x are uncorrelated and show no linear association, then the least squares fitted line is "horizontal," and identical to $\bar{y}$, so that $SSR = 0$ and $R^2 = 0$. When $0 < R^2 < 1$, it is interpreted as "the proportion of the variation in y about its mean that is explained by the regression model."

4.2.1 Correlation Analysis

In Appendix B.1.5, we discuss the **covariance** and **correlation** between two random variables x and y. The correlation coefficient ρ_{xy} between x and y is defined in (B.21) as

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.13}$$

In Appendix B, we did not discuss *estimating* the correlation coefficient. We will do so now to develop a useful relationship between the sample correlation coefficient and R^2.

Given a sample of data pairs $(x_i, y_i), i = 1, \ldots, N$, the sample correlation coefficient is obtained by replacing the covariance and standard deviations in (4.13) by their sample analogs:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) / (N - 1)$$

$$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (N - 1)}$$

$$s_y = \sqrt{\sum (y_i - \bar{y})^2 / (N - 1)}$$

The sample correlation coefficient r_{xy} has a value between -1 and 1, and it measures the strength of the linear association between observed values of x and y.

4.2.2 Correlation Analysis and R^2

There are two interesting relationships between R^2 and r_{xy} in the simple linear regression model.

1. The first is that $r_{xy}^2 = R^2$. That is, the square of the sample correlation coefficient between the sample data values x_i and y_i is algebraically equal to R^2 in a simple regression model. Intuitively, this relationship makes sense: r_{xy}^2 falls between zero and one and measures the strength of the linear association between x and y. This interpretation is not far from that of R^2: the proportion of variation in y about its mean explained by x in the linear regression model.
2. The second, and more important, relation is that R^2 can also be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = b_1 + b_2 x_i$. That is, $R^2 = r_{y\hat{y}}^2$. As such, it measures the linear association, or goodness-of-fit, between the sample data and their predicted values. Consequently, R^2 is sometimes called a measure of "goodness-of-fit." This result is valid not only in simple regression models but also in multiple regression models

that we introduce in Chapter 5. Furthermore, as you will see in Section 4.4, the concept of obtaining a goodness-of-fit measure by predicting y as well as possible and finding the squared correlation coefficient between this prediction and the sample values of y can be extended to situations in which the usual R^2 does not strictly apply.

EXAMPLE 4.2 | Goodness-of-Fit in the Food Expenditure Model

Look at the food expenditure example, Example 2.4, and in particular, the data scatter and fitted regression line in Figure 2.8, and the computer output in Figure 2.9. Go ahead. I will wait until you get back. The question we would like to answer is "How well does our model fit the data?" To compute R^2, we can use the sums of squares

$$SST = \sum(y_i - \bar{y})^2 = 495132.160$$

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum \hat{e}_i^2 = 304505.176$$

Then

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{304505.176}{495132.160} = 0.385$$

We conclude that 38.5% of the variation in food expenditure (about its sample mean) is explained by our regression model, which uses only income as an explanatory variable. Is this a good R^2? We would argue that such a question is not useful. Although finding and reporting R^2 provides information about the relative magnitudes of the different sources of variation, debates about whether a particular R^2 is "large enough" are not particularly constructive. Microeconomic household behavior is very difficult to explain. With cross-sectional data, R^2 values from 0.10 to 0.40 are very common even with much larger regression models. Macroeconomic analyses using time-series data, which often trend together smoothly over time, routinely report R^2 values of 0.90 and higher. You should *not* evaluate the quality of the model based only on how well it predicts the sample data used to construct the estimates. To evaluate the model, it is as important to consider factors such as the signs and magnitudes of the estimates, their statistical and economic significance, the precision of their estimation, and the ability of the fitted model to predict values of the dependent variable that were not in the estimation sample. Other model diagnostic issues will be discussed in the following section.

Correlation analysis leads to the same conclusions and numbers, but it is worthwhile to consider this approach in more detail. The sample correlation between the y and x sample values is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{478.75}{(6.848)(112.675)} = 0.62$$

The correlation is positive, indicating a positive association between food expenditure and income. The sample correlation measures the strength of the linear association, with a maximum value of 1. The value $r_{xy} = 0.62$ indicates a non-negligible but less than perfect fit. As expected $r_{xy}^2 = 0.62^2 = 0.385 = R^2$.

EXAMPLE 4.3 | Reporting Regression Results

In any paper where you write the results of a simple regression, with only one explanatory variable, these results can be presented quite simply. The key ingredients are the coefficient estimates, the standard errors (or t-values), an indication of statistical significance, and R^2. Also, when communicating regression results, avoid using symbols like x and y. Use abbreviations for the variables that are readily interpreted, defining the variables precisely in a separate section of the report. For the food expenditure example, we might have the variable definitions:

FOOD_EXP = weekly food expenditure by a household of size 3, in dollars

INCOME = weekly household income, in $100 units

Then the estimated equation results are as follows:

$$\begin{array}{lccc} FOOD\_EXP & = & 83.42 + 10.21\ INCOME & R^2 = 0.385 \\ (\text{se}) & & (43.41)^{*} \quad (2.09)^{***} & \end{array}$$

Report the standard errors below the estimated coefficients. The reason for showing the standard errors is that an approximate 95% interval estimate (if the degrees of freedom $N - 2$ are greater than 30) is $b_k \pm 2(\text{se})$. If desired, the reader may divide the estimate by the standard error to obtain the value of the t-statistic for testing a zero null hypothesis. Furthermore, testing other hypotheses is facilitated by having the standard error present. To test the null hypothesis $H_0: \beta_2 = 8.0$, we can quickly construct the t-statistic $t = [(10.21 - 8)/2.09]$ and proceed with the steps of the test procedure.

Asterisks are often used to show the reader the statistically significant (i.e., significantly different from zero using a two-tail test) coefficients, with explanations in a table footnote:

* indicates significant at the 10% level

** indicates significant at the 5% level

*** indicates significant at the 1% level

The asterisks are assigned by checking the p-values from the computer output, as shown in Figure 2.9.

4.3 Modeling Issues

4.3.1 The Effects of Scaling the Data

Data we obtain are not always in a convenient form for presentation in a table or use in a regression analysis. When the *scale* of the data is not convenient, it can be altered without changing any of the real underlying relationships between variables. For example, the real personal consumption in the United States, as of the second quarter of 2015, was \$12228.4 *billion* annually. That is, \$12,228,400,000,000 written out. While we *could* use the long form of the number in a table or in a regression analysis, there is no advantage to doing so. By choosing the units of measurement to be "billions of dollars," we have taken a long number and made it comprehensible. What are the effects of scaling the variables in a regression model?

Consider the food expenditure model. In Table 2.1 we report weekly expenditures in *dollars* but we report income in \$100 units, so a weekly income of \$2,000 is reported as $x = 20$. Why did we scale the data in this way? If we had estimated the regression using income in dollars, the results would have been

$$\begin{aligned} FOOD\_EXP &= 83.42 + 0.1021\ INCOME(\$) \quad R^2 = 0.385 \\ (\text{se}) \qquad &\quad (43.41)^{*}\ (0.0209)^{***} \end{aligned}$$

There are two changes. First, the estimated coefficient of income is now 0.1021. The interpretation is "If weekly household income increases by \$1 then we estimate that weekly food expenditure will increase by about 10 cents." There is nothing mathematically wrong with this, but it leads to a discussion of changes that are so small as to seem irrelevant. An increase in income of \$100 leads to an estimated increase in food expenditure of \$10.21, as before, but these magnitudes are more easily discussed.

The other change that occurs in the regression results when income is in dollars is that the standard error becomes smaller, by a factor of 100. Since the estimated coefficient is smaller by a factor of 100 also, this leaves the t-statistic and all other results unchanged.

Such a change in the units of measurement is called *scaling the data*. The choice of the scale is made by the researcher to make interpretation meaningful and convenient. The choice of the scale does not affect the measurement of the underlying relationship, but it does affect the interpretation of the coefficient estimates and some summary measures. Let us list the possibilities:

1. **Changing the scale of x**: In the linear regression model $y = \beta_1 + \beta_2 x + e$, suppose we change the units of measurement of the explanatory variable x by dividing it by a constant c. In order to keep intact the equality of the left- and right-hand sides, the coefficient of x must be multiplied by c. That is, $y = \beta_1 + \beta_2 x + e = \beta_1 + (c\beta_2)(x/c) + e = \beta_1 + \beta_2^* x^* + e$, where $\beta_2^* = c\beta_2$ and $x^* = x/c$. For example, if x is measured in dollars, and $c = 100$, then x^* is measured in hundreds of dollars. Then β_2^* measures the expected change in y given a \$100 increase in x, and β_2^* is 100 times larger than β_2. When the scale of x is altered, the only other change occurs in the standard error of the regression coefficient, but it changes by the same multiplicative factor as the coefficient, so that their ratio, the t-statistic, is unaffected. All other regression statistics are unchanged.

2. **Changing the scale of *y***: If we change the units of measurement of y, but not x, then all the coefficients must change in order for the equation to remain valid. That is, $y/c = (\beta_1/c) + (\beta_2/c)x + (e/c)$ or $y^* = \beta_1^* + \beta_2^* x + e^*$. In this rescaled model, β_2^* measures the change we expect in y^* given a 1-unit change in x. Because the error term is scaled in this process, the least squares residuals will also be scaled. This will affect the standard errors of the regression coefficients, but it will not affect t-statistics or R^2.
3. If the scale of y and the scale of x are changed by the same factor, then there will be no change in the reported regression results for b_2, but the estimated intercept and residuals will change; t-statistics and R^2 are unaffected. The interpretation of the parameters is made relative to the new units of measurement.

4.3.2 Choosing a Functional Form

In our ongoing example, we have assumed that the mean household food expenditure is a linear function of household income. That is, we assumed the underlying economic relationship to be $E(y|\mathbf{x}) = \beta_1 + \beta_2 x$, which implies that there is a linear, straight-line relationship between $E(y|\mathbf{x})$ and x. Why did we do that? Although the world is not "linear," a straight line is a good approximation to many nonlinear or curved relationships over narrow ranges. Moreover, in your principles of economics classes, you may have begun with straight lines for supply, demand, and consumption functions, and we wanted to ease you into the more "artistic" aspects of econometrics.

The starting point in all econometric analyses is economic theory. What does economics really say about the relation between food expenditure and income, holding all else constant? We expect there to be a positive relationship between these variables because food is a normal good. But nothing says the relationship must be a straight line. In fact, we do *not* expect that as household income rises, food expenditures will continue to rise indefinitely at the same constant rate. Instead, as income rises, we expect food expenditures to rise, but we expect such expenditures to increase at a decreasing rate. This is a phrase that is used many times in economics classes. What it means graphically is that there is not a straight-line relationship between the two variables. For a curvilinear relationship like that in Figure 4.4, the **marginal effect** of a change in the explanatory variable is measured by the slope of the tangent to the curve at a particular point. The marginal effect of a change in x is greater at the point (x_1, y_1) than it is at the point (x_2, y_2). As x increases, the value of y increases, but the slope is becoming smaller. This is the meaning of "increasing at a decreasing rate." In the economic context of the food expenditure model, the marginal propensity to spend on food is greater at lower incomes, and as income increases the marginal propensity to spend on food declines.

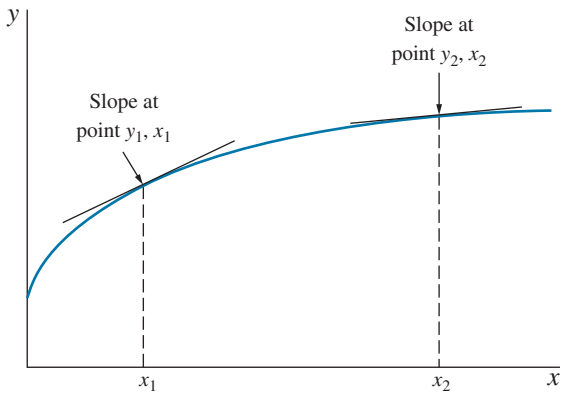


FIGURE 4.4 **A nonlinear relationship between food expenditure and income.**

The simple linear regression model is much more flexible than it appears at first glance. By *transforming* the variables y and x we can represent many curved, nonlinear relationships and still use the linear regression model. In Section 2.8, we introduced the idea of using **quadratic** and **log-linear** functional forms. In this and subsequent sections, we introduce you to an array of other possibilities and give some examples.

Choosing an algebraic form for the relationship means choosing *transformations* of the original variables. This is not an easy process, and it requires good analytic geometry skills and some experience. It may *not* come to you easily. The variable transformations that we begin with are as follows:

1. Power: If x is a variable, then x^p means raising the variable to the power p; examples are quadratic $\left(x^2\right)$ and cubic $\left(x^3\right)$ transformations.
2. The natural logarithm: If x is a variable, then its natural logarithm is $\ln(x)$.

Using just these two algebraic transformations, there are amazing varieties of "shapes" that we can represent, as shown in Figure 4.5.

A difficulty introduced when transforming variables is that regression result interpretations change. For each different functional form, shown in Table 4.1, the expressions for both the slope and elasticity change from the **linear relationship** case. This is so because the variables are related nonlinearly. What this means for the practicing economist is that great attention must be given to

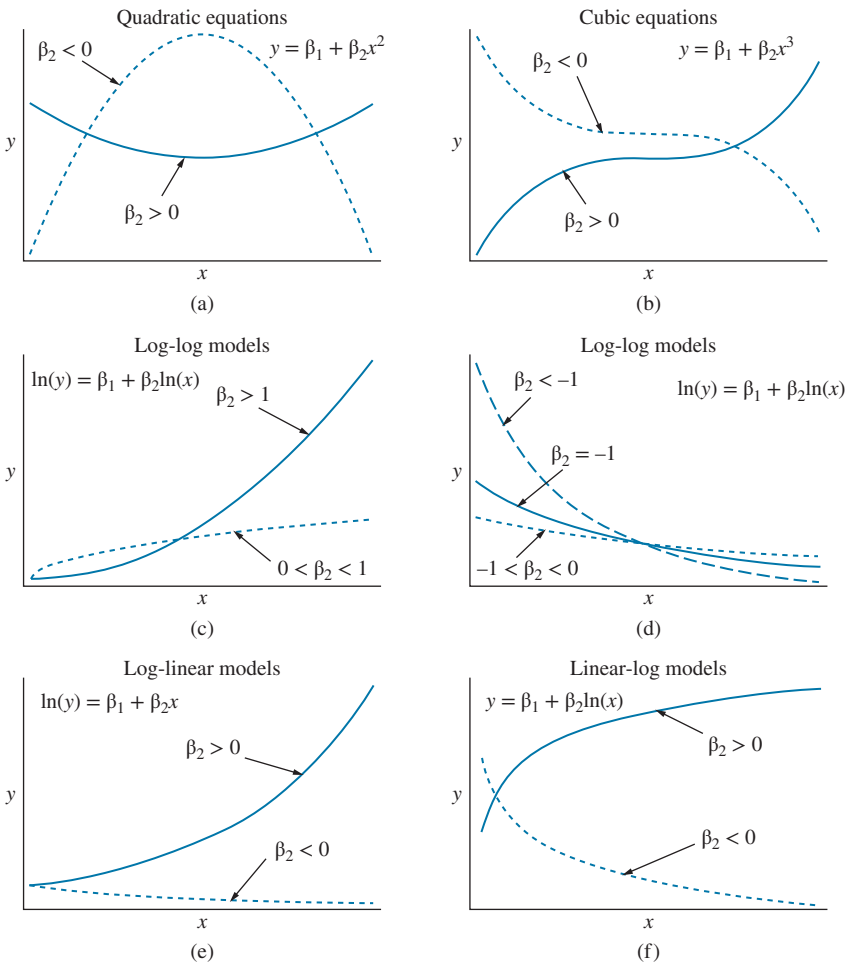


FIGURE 4.5 **Alternative functional forms.**

TABLE 4.1 **Some Useful Functions, Their Derivatives, Elasticities, and Other Interpretation**

Name	Function	Slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-linear	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1-unit change in x leads to (approximately) a $100\beta_2$% change in y		
Linear-log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

result interpretation whenever variables are transformed. Because you may be less familiar with logarithmic transformations, let us summarize the interpretation in three possible configurations.

1. In the **log-log model**, both the dependent and independent variables are transformed by the "natural" logarithm. The model is $\ln(y) = \beta_1 + \beta_2 \ln(x)$. In order to use this model, both y and x must be greater than zero because the logarithm is defined only for positive numbers. The parameter β_2 is the elasticity of y with respect to x. Referring to Figure 4.5, you can see why economists use the constant elasticity, log-log model specification so frequently. In panel (c), if $\beta_2 > 1$, the relation could depict a supply curve, or if $0 < \beta_2 < 1$, a production relation. In panel (d), if $\beta_2 < 0$, it could represent a demand curve. In each case, interpretation is convenient because the elasticity is constant. An example is given in Section 4.6.
2. In the **log-linear model** $\ln(y) = \beta_1 + \beta_2 x$, only the dependent variable is transformed by the logarithm. The dependent variable must be greater than zero to use this form. In this model, a 1-unit increase in x leads to (approximately) a $100\beta_2$% change in y. The log-linear form is common; it was introduced in Sections 2.8.3–2.8.4 and will be further discussed in Section 4.5. Note its possible shapes in Figure 4.5(e). If $\beta_2 > 0$, the function increases at an increasing rate; its slope is larger for larger values of y. If $\beta_2 < 0$, the function decreases, but at a decreasing rate.
3. In the **linear-log model** $y = \beta_1 + \beta_2 \ln(x)$ the variable x is transformed by the natural logarithm. See Figure 4.5(f). We can say that a 1% increase in x leads to a $\beta_2/100$-*unit* change in y. An example of this functional form is given in the following section.

Remark

Our plan for the remainder of this chapter is to consider several examples of the uses of alternative functional forms. In the following section we use the linear-log functional form with the food expenditure data. Then we take a brief detour into some diagnostic measures for data and model adequacy based on the least squares residuals. After discussing the diagnostic tools we give examples of polynomial equations, log-linear equations, and log-log equations.

4.3.3 A Linear-Log Food Expenditure Model

Suppose that in the food expenditure model, we wish to choose a functional form that is consistent with Figure 4.4. One option is the linear-log functional form. A linear-log equation has a

linear, untransformed term on the left-hand side and a logarithmic term on the right-hand side, or $y = \beta_1 + \beta_2 \ln(x)$. Because of the logarithm, this function requires $x > 0$. It is an increasing or decreasing function, depending on the sign of β_2. Using Derivative Rule 8, Appendix A, the slope of the function is β_2/x, so that as x increases, the slope decreases in absolute magnitude. If $\beta_2 > 0$, then the function increases at a decreasing rate. If $\beta_2 < 0$, then the function decreases at a decreasing rate. The function shapes are depicted in Figure 4.5(f). The elasticity of y with respect to x in this model is $\varepsilon = slope \times x/y = \beta_2/y$.

There is a convenient interpretation using approximations to changes in logarithms. Consider a small increase in x from x_0 to x_1. Then $y_0 = \beta_1 + \beta_2 \ln(x_0)$ and $y_1 = \beta_1 + \beta_2 \ln(x_1)$. Subtracting the former from the latter, and using the approximation developed in Appendix A, equation (A.3), gives

$$\begin{aligned} \Delta y = y_1 - y_0 &= \beta_2 \left[\ln(x_1) - \ln(x_0) \right] \\ &= \frac{\beta_2}{100} \times 100 \left[\ln(x_1) - \ln(x_0) \right] \\ &\cong \frac{\beta_2}{100} (\%\Delta x) \end{aligned}$$

The change in y, represented in its units of measure, is approximately $\beta_2/100$ times the percentage change in x.

EXAMPLE 4.4 | Using the Linear-Log Model for Food Expenditure

Using a linear-log equation for the food expenditure relation results in the regression model

$$FOOD\_EXP = \beta_1 + \beta_2 \ln(INCOME) + e$$

For $\beta_2 > 0$ this function is increasing but at a decreasing rate. As *INCOME* increases the slope $\beta_2/INCOME$ decreases. In this context, the slope is the marginal propensity to spend on food from additional income. Similarly, the elasticity, $\beta_2/FOOD\_EXP$, becomes smaller for larger levels of food expenditure. These results are consistent with the idea that at high incomes, and large food expenditures, the effect of an increase in income on food expenditure is small.

The estimated linear-log model using the food expenditure data is

$$\begin{aligned} \widehat{FOOD\_EXP} &= -97.19 + 132.17 \ln(INCOME) \quad R^2 = 0.357 \\ (se) \qquad & (84.24) \quad (28.80)^{***} \end{aligned} \tag{4.14}$$

The fitted model is shown in Figure 4.6.

As anticipated, the fitted function is not a straight line. The fitted linear-log model is consistent with our theoretical model that anticipates declining marginal propensity to spend additional income on food. For a household with \$1,000 weekly income, we estimate that the household will spend an additional \$13.22 on food from an additional

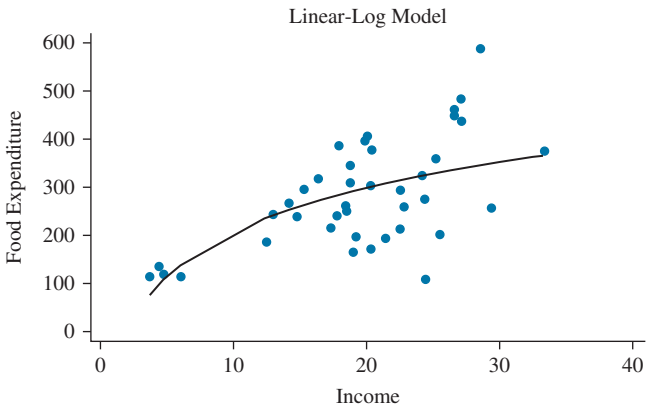


FIGURE 4.6 **The fitted linear-log model.**

$100 income, whereas we estimate that a household with $2,000 per week income will spend an additional $6.61 from an additional $100 income. The marginal effect of income on food expenditure is smaller at higher levels of income. This is a change from the linear, straight-line relationship we originally estimated, in which the marginal effect of a change in income of $100 was $10.21 for all levels of income.

Alternatively, we can say that a 1% increase in income will increase food expenditure by approximately $1.32 per week or that a 10% increase in income will increase food expenditure by approximately $13.22. Although this interpretation is conveniently simple to state, the diminishing marginal effect of income on food expenditure is somewhat disguised, though still implied. At $1,000 per week income, a 10% increase is $100, while at $2,000 income a 10% increase is $200. At higher levels of income, a larger dollar increase in income is required to elicit an additional $13.22 expenditure on food.

In terms of how well the model fits the data, we see that $R^2 = 0.357$ for the linear-log model, as compared to $R^2 = 0.385$ for the linear, straight-line relationship. Since these two models have the same dependent variable, *FOOD_EXP*, and each model has a single explanatory variable, a comparison of R^2 values is valid. However, there is a very small difference in the fit of the two models, and in any case, a model should not be chosen only on the basis of model fit with R^2 as the criterion.

Remark

Given alternative models that involve different transformations of the dependent and independent variables, and some of which have similar shapes, what are some guidelines for choosing a functional form?

1. Choose a shape that is consistent with what economic theory tells us about the relationship.
2. Choose a shape that is sufficiently flexible to "fit" the data.
3. Choose a shape so that assumptions SR1–SR6 are satisfied, ensuring that the least squares estimators have the desirable properties described in Chapters 2 and 3.

Although these objectives are easily stated, the reality of model building is much more difficult. You must recognize that we **never** know the "true" functional relationship between economic variables; also, the functional form that we select, no matter how elegant, is only an approximation. Our job is to choose a functional form that satisfactorily meets the three objectives stated above.

4.3.4 Using Diagnostic Residual Plots

When specifying a regression model, we may inadvertently choose an inadequate or incorrect functional form. Even if the functional form is adequate, one or more of the regression model assumptions may not hold. There are two primary methods for detecting such errors. First, examine the regression results. Finding an incorrect sign or a theoretically important variable that is not statistically significant may indicate a problem. Second, evidence of specification errors can reveal themselves in an analysis of the least squares residuals. We should ask whether there is any evidence that assumptions SR3 (homoskedasticity), SR4 (no serial correlation), and SR6 (normality) are violated. Usually, heteroskedasticity might be suspected in cross-sectional data analysis, and serial correlation is a potential time-series problem. In both cases, diagnostic tools focus on the least squares residuals. In Chapters 8 and 9, we will provide formal tests for homoskedasticity and serial correlation. In addition to formal tests, residual plots of all types are useful as diagnostic tools. In this section, residual analysis reveals potential heteroskedasticity and serial correlation problems and also flawed choices of functional forms.

We show a variety of residual plots in Figure 4.7. If there are no violations of the assumptions, then a plot of the least squares residuals versus x, y, or the fitted value of y, $\hat{y}$, should reveal no patterns. Figure 4.7(a) is an example of a random scatter.

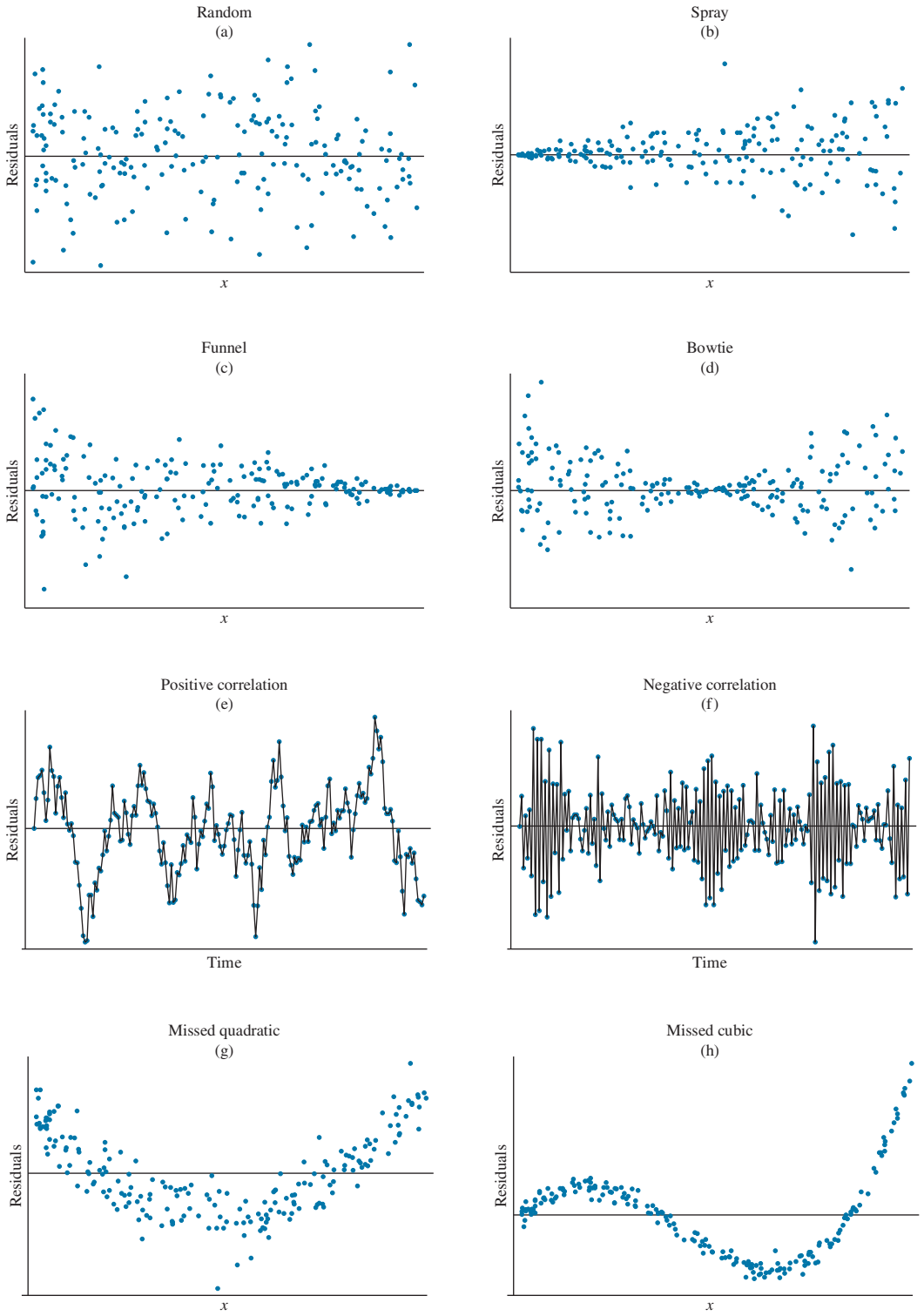


FIGURE 4.7 **Residual patterns.**

Figures 4.7(b)–(d) show patterns associated with heteroskedasticity. Figure 4.7(b) has a "spray-shaped" residual pattern that is consistent with the variance of the error term increasing as x-values increase; Figure 4.7(c) has a "funnel-shaped" residual pattern that is consistent with the variance of the error term decreasing as x-values increase; and Figure 4.7(d) has a "bow-tie" residual pattern that is consistent with the variance of the error term decreasing and then increasing as x-values increase.

Figure 4.7(e) shows a typical pattern produced with time-series regression when the error terms display a positive correlation, $\text{corr}(e_t, e_{t-1}) > 0$. Note that there are sequences of positive residuals followed by sequences of negative residuals, and so on. If assumption SR4 holds there should be no such sign patterns. Figure 4.7(f) shows a typical pattern produced with time-series regression when the error terms display a negative correlation, $\text{corr}(e_t, e_{t-1}) < 0$. In this case, each positive residual tends to be followed by a negative residual, which is then followed by a positive residual and so on. The sequence of residuals tends to alternate in sign.

If the relationship between y and x is curvilinear, such as a U-shaped quadratic function, like an average cost function, and we mistakenly assume that the relationship is linear, then the least squares residuals may show a U-shape like in Figure 4.7(g). If the relationship between y and x is curvilinear, such as a cubic function, like a total cost function, and we mistakenly assume that the relationship is linear, then the least squares residuals may show a serpentine shape like Figure 4.7(h).

The bottom line is that when least squares residuals are plotted against another variable there should be no patterns evident. Patterns of the sorts shown in Figure 4.7, except for panel (a), indicate that there may be some violation of assumptions and/or incorrect model specification.

EXAMPLE 4.5 | Heteroskedasticity in the Food Expenditure Model

The least squares residuals from the linear-log food expenditure model in (4.14) are plotted in Figure 4.8. These exhibit an expanding variation pattern with more variation in the residuals as *INCOME* becomes larger, which may suggest heteroskedastic errors. A similar residual plot is implied by Figure 2.8.

We must conclude that at this point we do not have a satisfactory model for the food expenditure data. The linear and linear-log models have different shapes and different implied marginal effects. The two models fit the data equally well, but both models exhibit least squares residual patterns consistent with heteroskedastic errors. This example will be considered further in Chapter 8.

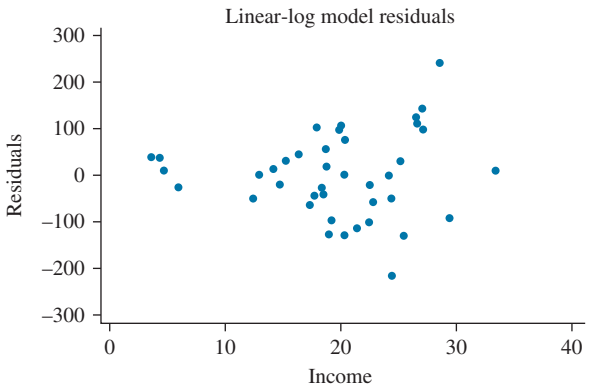


FIGURE 4.8 **Residuals from linear-log food expenditure model.**

4.3.5 Are the Regression Errors Normally Distributed?

Recall that hypothesis tests and interval estimates for the coefficients rely on SR6 assumption, that given $\mathbf{x}$, the errors, and hence the dependent variable y, are normally distributed. Though our tests and confidence intervals are valid in large samples whether the data are normally distributed or not, it is nevertheless desirable to have a model in which the regression errors are normally distributed, so that we do not have to rely on large sample approximations. If the errors are not normally distributed, we might be able to improve our model by considering an alternative functional form or transforming the dependent variable. As noted in the last "Remark," when choosing

a functional form, one of the criteria we might examine is whether a model specification satisfies regression assumptions, and in particular, whether it leads to errors that are normally distributed (SR6). How do we check out the assumption of normally distributed errors?

We cannot observe the true random errors, so we must base our analysis of their normality on the least squares residuals, $\hat{e}_i = y_i - \hat{y}_i$. Substituting for y_i and $\hat{y}_i$, we obtain

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + e_i - (b_1 + b_2 x_i) \\ &= (\beta_1 - b_1) + (\beta_2 - b_2)x_i + e_i \\ &= e_i - (b_1 - \beta_1) - (b_2 - \beta_2)x_i\end{aligned}$$

In large samples, $(b_1 - \beta_1)$ and $(b_2 - \beta_2)$ will tend toward zero because the least squares estimators are unbiased and have variances that approach zero as $N \to \infty$. Consequently, in large samples, the difference $\hat{e}_i - e_i$ is close to zero, so that these two random variables are essentially the same and thus have the same distribution.

A histogram of the least squares residuals gives us a graphical representation of the empirical distribution.

EXAMPLE 4.6 | Testing Normality in the Food Expenditure Model

The relevant EViews output for the food expenditure example, using the linear relationship with no transformation of the variables, appears in Figure 4.9. What does this histogram tell us? First, notice that it is centered at zero. This is not surprising because the mean of the least squares residuals is always zero if the model contains an intercept, as shown in Appendix 4B. Second, it seems symmetrical, but there are some large gaps, and it does not really appear bell shaped. However, merely checking the shape of the histogram, especially when the number of observations is relatively small, is not a statistical "test."

There are many tests for normality. The **Jarque–Bera test** for normality is valid in large samples. It is based on two measures, **skewness** and **kurtosis**. In the present context, **skewness** refers to how symmetric the residuals are around zero. Perfectly symmetric residuals will have a skewness of zero. The skewness value for the food expenditure residuals is −0.097. **Kurtosis** refers to the "peakedness" of the distribution. For a normal distribution, the kurtosis value is 3. For more on skewness and kurtosis, see Appendices B.1.2 and C.4.2. From Figure 4.9, we see that the food expenditure residuals have a kurtosis of 2.99. The skewness and kurtosis values are close to the values for the normal distribution. So, the question we have to ask is whether 2.99 is sufficiently different from 3, and −0.097 is sufficiently different from zero, to conclude that the residuals are not normally distributed. The Jarque–Bera statistic is given by

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

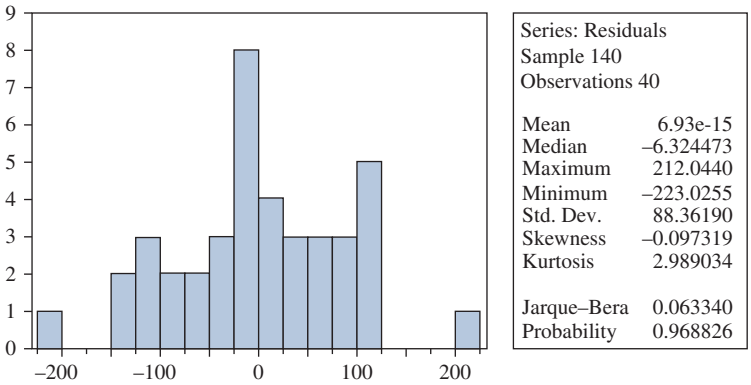


FIGURE 4.9 **EViews output: residuals histogram and summary statistics for food expenditure example.**

where N is the sample size, S is skewness, and K is kurtosis. Thus, large values of the skewness, and/or values of kurtosis quite different from 3, will lead to a large value of the Jarque–Bera statistic. When the residuals are normally distributed, the Jarque–Bera statistic has a chi-squared distribution with two degrees of freedom. We reject the hypothesis of normally distributed errors if a calculated value of the statistic exceeds a critical value selected from the chi-squared distribution with two degrees of freedom. Using Statistical Table 3, the 5% critical value from a χ^2-distribution with two degrees of freedom is 5.99, and the 1% critical value is 9.21.

Applying these ideas to the food expenditure example, we have

$$JB = \frac{40}{6}\left((-0.097)^2 + \frac{(2.99-3)^2}{4}\right) = 0.063$$

Because $0.063 < 5.99$, there is insufficient evidence from the residuals to conclude that the normal distribution assumption is unreasonable at the 5% level of significance. The same conclusion could have been reached by examining the p-value. The p-value appears in Figure 4.9 described as "Probability." Thus, we also fail to reject the null hypothesis on the grounds that $0.9688 > 0.05$.

For the linear-log model of food expenditure reported in Example 4.4, the Jarque–Bera test statistic value is 0.1999 with a p-value of 0.9049. We cannot reject the null hypothesis that the regression errors are normally distributed, and this criterion does not help us choose between the linear and linear-log functional forms for the food expenditure model.

In these examples, we should remember that the Jarque–Bera test is strictly valid only in large samples. Applying tests that are valid in large samples to smaller samples, such as $N = 40$, is not uncommon in applied work. However, we should remember in such applications that we should not give great weight to the test significance or nonsignificance.

4.3.6 Identifying Influential Observations

One worry in data analysis is that we may have some unusual and/or **influential observations**. Sometimes, these are termed "outliers." If an unusual observation is the result of a data error, then we should correct it. If an unusual observation is not the result of a data error, then understanding how it came about, the story behind it, can be informative. One way to detect whether an observation is influential is to delete it and reestimate the model, comparing the results to the original results based on the full sample. This "delete-one" strategy can help detect the influence of the observation on the estimated coefficients and the model's predictions. It can also help us identify unusual observations.

The delete-one strategy begins with the least squares parameter estimates based on the sample with the ith observation deleted. Denote these as $b_1(i)$ and $b_2(i)$. Let $\hat{\sigma}^2(i)$ be the delete-one estimated error variance. The residual $\hat{e}(i) = y_i - \left[b_1(i) + b_2(i)\,x_i\right]$ is the actual value of y for the ith observation, y_i, minus the fitted value that uses estimates from the sample with the ith observation deleted. It is the forecast error (4.3) with y_i taking the place of y_0 and x_i taking the value of x_0 and using the estimates $b_1(i)$ and $b_2(i)$. Modifying the variance of the forecast error (4.4), we obtain the variance of $\hat{e}(i)$ (and its estimator) as

$$\widehat{\text{var}}\left[\hat{e}(i)|\mathbf{x}\right] = \hat{\sigma}^2(i)\left[1 + \frac{1}{(N-1)} + \frac{\left(x_i - \bar{x}(i)\right)^2}{\sum_{j\neq i}\left(x_j - \bar{x}(i)\right)^2}\right]$$

where $\bar{x}(i)$ is the delete-one sample mean of the x-values. The ratio

$$\hat{e}_i^{\text{stu}} = \frac{\hat{e}(i)}{\left\{\widehat{\text{var}}\left[\hat{e}(i)|\mathbf{x}\right]\right\}^{1/2}}$$

is called a **studentized residual**. It is the standardized residual based on the delete-one sample. The rule of thumb is to calculate these values and compare their values to ± 2, which is roughly

a 95% interval estimate. If the studentized residual falls outside the interval, then the observation is worth examining because it is "unusually" large.

After considerable algebra, the studentized residual can also be written as

$$\hat{e}_i^{\text{stu}} = \frac{\hat{e}_i}{\hat{\sigma}(i)\left(1 - h_i\right)^{1/2}}$$

where

$$h_i = \frac{1}{N} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}$$

The term h_i is called the **leverage** of the ith observation, with values $0 \le h_i \le 1$. If the leverage value is high, then the value of the studentized residual is inflated. The second component of h_i is $\left(x_i - \bar{x}\right)^2 / \sum\left(x_i - \bar{x}\right)^2$. Recall that the sample variance of the x_i-values is estimated by $s_x^2 = \sum\left(x_i - \bar{x}\right)^2 / (N-1)$ so that $\sum\left(x_i - \bar{x}\right)^2$ is a measure of the total variation in the sample x_i-values about their mean. If one observation's contribution $\left(x_i - \bar{x}\right)^2$ to the total is large, then that observation may have a strong effect on the least squares estimates and fitted values. The sum of the leverage terms h_i is K, the number of parameters in the regression model. Thus, the average value in the simple regression model is $\bar{h} = K/N = 2/N$. When checking data, it is a common rule of thumb to examine observations with leverage greater than two or three times the average.

Another measure of the influence of a single observation on the least squares estimates is called **DFBETAS**. For the slope estimate in the simple regression model, we calculate

$$\text{DFBETAS}_{2i} = \frac{b_2 - b_2(i)}{\hat{\sigma}(i) \Big/ \sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}}$$

The effect of the ith observation on the slope estimate is measured by the change in the coefficient estimate by dropping the ith observation and then standardizing. The magnitude of DFBETAS_{2i} will be larger when leverage is larger and/or the studentized residual is larger. A common rule of thumb for identifying influential observations in the simple regression model is $\left|\text{DFBETAS}_{2i}\right| > 2/\sqrt{N}$.

The effect of the ith observation on the fitted value from the least squares regression is again a measurement using the delete-one approach. Let $\hat{y}_i = b_1 + b_2 x_i$ and $\hat{y}(i) = b_1(i) + b_2(i)\, x_i$ with $\hat{y}_{(i)}$ being the fitted value using parameter estimates from the delete-one sample. The measure called **DFFITS** is

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}(i)}{\hat{\sigma}(i)\, h_i^{1/2}} = \left(\frac{h_i}{1 - h_i}\right)^{1/2} \hat{e}_i^{\text{stu}}$$

This measure will be larger when leverage is larger and/or the studentized residual is larger. A rule of thumb to identify unusual observations is $\left|\text{DFFITS}_i\right| > 2(K/N)^{1/2}$ or $\left|\text{DFFITS}_i\right| > 3(K/N)^{1/2}$ where $K = 2$ is the number of parameters in the simple regression model.

These constructs may look difficult to compute, but modern software usually computes some or all of these measures. We are **not** suggesting that you toss out unusual observations. If these measures lead you to locate an observation with an error, you can try to fix it. By looking at unusual observations, ones that have a high leverage, a large studentized residual, a large DFBETAS, or a large DFFITS, you may learn something about which data characteristics are important. All data analysts should examine their data, and these tools may help organize such an examination.

EXAMPLE 4.7 | Influential Observations in the Food Expenditure Data

Examining the influential observation measures for the food expenditure data, using a linear relationship and no transformations of the variables, reveals few real surprises. First the leverage values have the average $\bar{h} = 2/40 = 0.05$. Isolating observations with leverage more than twice the average, we have

obs	*h*	*FOOD_EXP*	*INCOME*
1	0.1635	115.22	3.69
2	0.1516	135.98	4.39
3	0.1457	119.34	4.75
4	0.1258	114.96	6.03
40	0.1291	375.73	33.4

The observations with the greatest leverage are those with the four lowest incomes and the highest income. The mean of *INCOME* is 19.6.

The observations with studentized residuals, *EHATSTU*, larger than two in absolute value are

obs	*EHATSTU*	*FOOD_EXP*	*INCOME*
31	−2.7504	109.71	24.42
38	2.6417	587.66	28.62

These two observations are interesting because the food expenditures for these two households are the minimum and maximum, despite both incomes being above the mean. In fact, the income for household 31 is the 75th percentile value, and the income for household 38 is the third largest. Thus, household 31 is spending significantly less on food than we would predict, and household 38 more than we would predict, based on income alone. These might be observations worth checking to ensure they are correct. In our case, they are.

The DFBETAS values greater than $2/\sqrt{N} = 0.3162$ in absolute value are

obs	*DFBETAS*	*FOOD_EXP*	*INCOME*
38	0.5773	587.66	28.62
39	−0.3539	257.95	29.40

Again household 38 has a relatively large influence on the least squares estimate of the slope. Household 39 shows up because it has the second highest income but spends less than the mean value (264.48) on food.

Finally, DFFITS values larger than $2(K/N)^{1/2} = 0.4472$ are as follows:

obs	*DFFITS*	*FOOD_HAT*	*FOOD_EXP*	*INCOME*
31	−0.5442	332.74	109.71	24.42
38	0.7216	375.62	587.66	28.62

The observations with a high influence of the least squares fitted values are the previously mentioned households 31 and 38, which also have large studentized residuals.

4.4 Polynomial Models

In Sections 2.8.1–2.8.2, we introduced the use of quadratic polynomials to capture curvilinear relationships. Economics students will have seen many average and marginal cost curves (U-shaped) and average and marginal product curves (inverted-U shaped) in their studies. Higher order polynomials, such as cubic equations, are used for total cost and total product curves. A familiar example to economics students is the total cost curve, shaped much like the solid curve in Figure 4.5(b). In this section, we review simplified quadratic and cubic equations and give an empirical example.

4.4.1 Quadratic and Cubic Equations

The general form of a quadratic equation $y = a_0 + a_1x + a_2x^2$ includes a constant term a_0, a linear term a_1x, and a squared term a_2x^2. Similarly, the general form of a cubic equation is $y = a_0 + a_1x + a_2x^2 + a_3x^3$. In Section 5.6, we consider multiple regression models using the

general forms of quadratic and cubic equations. For now, however, because we are working with "simple" regression models that include only one explanatory variable, we consider the simple quadratic and cubic forms, $y = \beta_1 + \beta_2 x^2$ and $y = \beta_1 + \beta_2 x^3$, respectively. The properties of the simple quadratic function are discussed in Section 2.8.1.

The simple cubic equation $y = \beta_1 + \beta_2 x^3$ has possible shapes shown in Figure 4.5(b). Using Derivative Rules 4 and 5 from Appendix A, the derivative, or slope, of this cubic equation is $dy/dx = 3\beta_2 x^2$. The slope of the curve is always positive if $\beta_2 > 0$, except when $x = 0$, yielding a direct relationship between y and x like the solid curve shown in Figure 4.5(b). If $\beta_2 < 0$, then the relationship is an inverse one like the dashed curve shown in Figure 4.5(b). The slope equation shows that the slope is zero only when $x = 0$. The term β_1 is the y-intercept. The elasticity of y with respect to x is $\varepsilon = slope \times x/y = 3\beta_2 x^2 \times x/y$. Both the slope and elasticity change along the curve.

EXAMPLE 4.8 | An Empirical Example of a Cubic Equation

Figure 4.10 is a plot of average wheat yield (in tonnes per hectare—a hectare is about 2.5 acres, and a tonne is a metric ton that is 1000 kg or 2205 lb—we are speaking Australian here!) for the Greenough Shire in Western Australia, against time. The observations are for the period 1950–1997, and time is measured using the values 1, 2, …, 48. These data can be found in the data file *wa_wheat*. Notice in Figure 4.10 that wheat yield fluctuates quite a bit, but overall, it tends to increase over time, and the increase is at an increasing rate, particularly toward the end of the time period. An increase in yield is expected because of technological improvements, such as the development of varieties of wheat that are higher yielding and more resistant to pests and diseases. Suppose that we are interested in measuring the effect of technological improvement on yield. Direct data on changes in technology are not available, but we can examine how wheat yield has changed over time as a consequence of changing technology. The equation of interest relates *YIELD* to *TIME*, where $TIME = 1, \ldots, 48$. One problem with the linear equation

$$YIELD_t = \beta_1 + \beta_2 TIME_t + e_t$$

is that it implies that yield increases at the same constant rate β_2, when, from Figure 4.10, we expect this rate to be increasing. The least squares fitted line (standard errors in parentheses) is

$$\widehat{YIELD}_t = 0.638 + 0.0210\, TIME_t \qquad R^2 = 0.649$$
$$\text{(se)} \qquad (0.064)\ (0.0022)$$

The residuals from this regression are plotted against time in Figure 4.11. Notice that there is a concentration of positive residuals at each end of the sample and a concentration of negative residuals in the middle. These concentrations are caused by the inability of a straight line to capture the fact that yield is increasing at an increasing rate. Compare the residual pattern in Figure 4.11 to Figures 4.7(g) and (h). What alternative can we try? Two possibilities are $TIME^2$ and $TIME^3$. It turns out that $TIME^3$ provides the better fit, and so we consider the functional form

$$YIELD_t = \beta_1 + \beta_2 TIME_t^3 + e_t$$

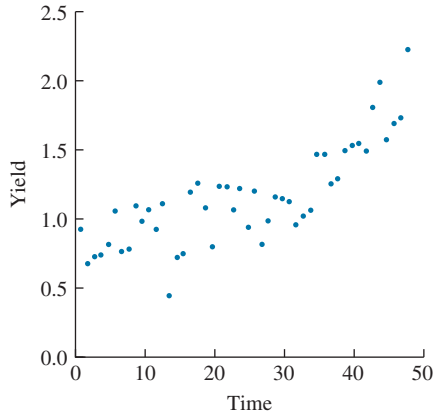


FIGURE 4.10 Scatter plot of wheat yield over time.

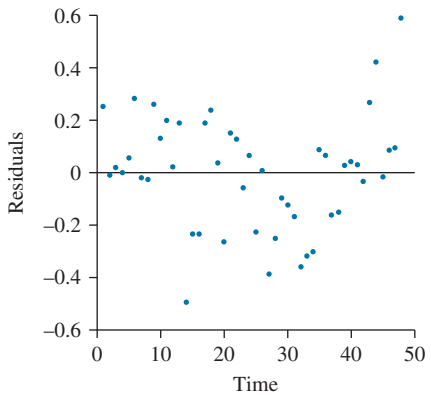


FIGURE 4.11 Residuals from a linear yield equation.

The slope of the expected yield function is $3\beta_2 TIME^2$. Thus, so long as the estimate of β_2 turns out to be positive, the function will be increasing. Furthermore, the slope is increasing as well. Thus, the function itself is "increasing at an increasing rate." Before estimating the cubic equation, note that the values of $TIME^3$ can get very large. This variable is a good candidate for scaling. If we define $TIMECUBE_t = (TIME_t/100)^3$, the estimated equation is

$$\widehat{YIELD}_t = 0.874 + 9.682\,TIMECUBE_t \qquad R^2 = 0.751$$
$$\text{(se)} \qquad (0.036)\ (0.822)$$

The residuals from this cubic equation are plotted in Figure 4.12. The predominance of positive residuals at the ends and negative residuals in the middle no longer exists. Furthermore, the R^2 value has increased from 0.649 to 0.751, indicating that the equation with *TIMECUBE* fits the data better than the one with just *TIME*. Both these equations have the same dependent variable and the same number of explanatory variables (only 1). In these circumstances, the R^2 can be used legitimately to compare goodness-of-fit.

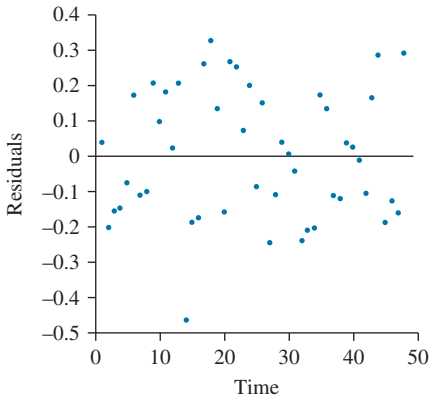


FIGURE 4.12 **Residuals from a cubic yield equation.**

What lessons have we learned from this example? First, a plot of the original dependent variable series y against the explanatory variable x is a useful starting point for deciding on a functional form in a simple regression model. Secondly, examining a plot of the residuals is a useful device for uncovering inadequacies in any chosen functional form. Runs of positive and/or negative residuals can suggest an alternative. In this example, with time-series data, plotting the residuals against time was informative. With cross-sectional data, using plots of residuals against both independent and dependent variables is recommended. Ideally, we will see no patterns, and the residual histogram and Jarque–Bera test will not rule out the assumption of normality. As we travel through the book, you will discover that patterns in the residuals, such as those shown in Figure 4.7, can also mean many other specification inadequacies, such as omitted variables, heteroskedasticity, and autocorrelation. Thus, as you become more knowledgeable and experienced, you should be careful to consider other options. For example, wheat yield in Western Australia is heavily influenced by rainfall. Inclusion of a rainfall variable might be an option worth considering. Also, it makes sense to include *TIME* and $TIME^2$ in addition to $TIME^3$. A further possibility is the constant growth rate model that we consider in the following section.

4.5 Log-Linear Models

Econometric models that employ natural logarithms are very common. We first introduced the log-linear model in Section 2.8.3. Logarithmic transformations are often used for variables that are monetary values, such as wages, salaries, income, prices, sales, and expenditures, and, in general, for variables that measure the "size" of something. These variables have the characteristic that they are positive and often have distributions that are positively skewed, with a long tail to the right. Figure P.2 in the Probability Primer is representative of the income distribution in the United States. In fact, the probability density function $f(x)$ shown is called the "log-normal" because $\ln(x)$ has a normal distribution. Because the transformation $\ln(x)$ has the effect of making larger values of x less extreme, $\ln(x)$ will often be closer to a normal distribution for variables of this kind. The **log-normal distribution** is discussed in Appendix B.3.9.

The log-linear model, $\ln(y) = \beta_1 + \beta_2 x$, has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side. Both its slope and elasticity change at each point and are the same sign as β_2. Using the antilogarithm, we obtain $\exp[\ln(y)] = y = \exp(\beta_1 + \beta_2 x)$, so that the log-linear function is an exponential function. The function requires $y > 0$. The slope at any point is $\beta_2 y$, which for $\beta_2 > 0$ means that the marginal effect increases for larger values of y. An economist might say that this function is increasing at an increasing rate. The shapes of the log-linear model are shown in Figure 4.5(e), and its derivative and elasticity given in Table 4.1. To make discussion relevant in a specific context, the slope can be evaluated at the sample mean $\bar{y}$, or the elasticity $\beta_2 x$ can be evaluated at the sample mean $\bar{x}$, or other interesting values can be chosen.

Using the properties of logarithms, we can obtain a useful approximation. Consider an increase in x from x_0 to x_1. The change in the log-linear model is from $\ln(y_0) = \beta_1 + \beta_2 x_0$ to $\ln(y_1) = \beta_1 + \beta_2 x_1$. Subtracting the first equation from the second gives $\ln(y_1) - \ln(y_0) = \beta_2(x_1 - x_0) = \beta_2 \Delta x$. Multiply by 100, and use the approximation introduced in Appendix A, equation (A.3) to obtain

$$100\left[\ln(y_1) - \ln(y_0)\right] \cong \%\Delta y = 100\beta_2(x_1 - x_0) = (100\beta_2) \times \Delta x$$

A 1-unit increase in x leads approximately to a $100\beta_2\%$ change in y.

In the following two examples, we apply the familiar concept of **compound interest** to derive a log-linear economic model for growth arising from technology, and a model explaining the relation between an individual's wage rate and their years of schooling. Recall the compound interest formula. If an investor deposits an initial amount V_0 (the principal amount) into an account that earns a rate of return r, then after t periods the value V of the account is $V_t = V_0(1 + r)^t$. For example, if $r = 0.10$, so that the rate of return is 10%, and if $V_0 = \$100$, after one period the account value is $V_1 = \$110$; after two periods, the account value is $V_2 = \$121$, and so on. The compound interest formula also explains the account growth from year to year. The accumulated value earns the rate r in each period so that $V_t = V_0(1 + r)^t = (1 + r)V_{t-1}$.

EXAMPLE 4.9 | A Growth Model

Earlier in this chapter, in Example 4.8, we considered an empirical example in which the production of wheat was tracked over time, with improvements in technology leading to wheat production increasing at an increasing rate. We observe wheat production in time periods $t = 1, \ldots, T$. Assume that in each period *YIELD* grows at the constant rate g due to technological progress. Let the *YIELD* at time $t = 0$, before the sample begins, be $YIELD_0$. This plays the role of the initial amount. Applying the compound interest formula we have $YIELD_t = YIELD_0(1 + g)^t$. Taking logarithms, we obtain

$$\begin{aligned}\ln(YIELD_t) &= \ln(YIELD_0) + [\ln(1 + g)] \times t \\ &= \beta_1 + \beta_2 t\end{aligned}$$

This is simply a log-linear model with dependent variable $\ln(YIELD_t)$ and explanatory variable t, or time. We expect growth to be positive, so that $\beta_2 > 0$, in which case the plot of *YIELD* against time looks like the upward-sloping curve in Figure 4.5(c), which closely resembles the scatter diagram in Figure 4.11.

Estimating the log-linear model for yield, we obtain

$$\begin{aligned}\widehat{\ln(YIELD_t)} &= -0.3434 + 0.0178t \\ (\text{se}) \qquad & (0.0584) \quad (0.0021)\end{aligned}$$

The estimated coefficient $b_2 = \widehat{\ln(1 + g)} = 0.0178$. Using the property that $\ln(1 + x) \cong x$ if x is small [see Appendix A, equation (A.4) and the discussion following it], we estimate that the growth rate in wheat yield is approximately $\hat{g} = 0.0178$, or about 1.78% per year, over the period of the data.

EXAMPLE 4.10 | A Wage Equation

The relationship between wages and education is a key relationship in labor economics (and, no doubt, in your mind). Suppose that the rate of return to an extra year of education is a constant r. Let $WAGE_0$ represent the wage of a person with no education. Applying the compound interest formula to the investment in human capital, we anticipate that the wage of a person with one year of education will be $WAGE_1 = WAGE_0(1 + r)$. A second year of education will compound the human capital so that $WAGE_2 = WAGE_1(1 + r) = WAGE_0(1 + r)^2$. In general, $WAGE = WAGE_0(1 + r)^{EDUC}$, where $EDUC$ is years of education. Taking logarithms, we have a relationship between $\ln(WAGE)$ and years of education ($EDUC$)

$$\begin{aligned}\ln(WAGE) &= \ln\left(WAGE_0\right) + \left[\ln(1 + r)\right] \times EDUC \\ &= \beta_1 + \beta_2 EDUC\end{aligned}$$

An additional year of education leads to an approximate $100\beta_2\%$ increase in wages.

Data on hourly wages, years of education, and other variables are in the file *cps5_small*. These data consist of 1200 observations from the May 2013 Current Population Survey (CPS). The CPS is a monthly survey of about 50000 households conducted in the United States by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years. Using these data, the estimated log-linear model is

$$\begin{aligned}\widehat{\ln(WAGE)} &= 1.5968 + 0.0988 \times EDUC \\ \text{(se)} \qquad & (0.0702) \quad (0.0048)\end{aligned}$$

We estimate that an additional year of education increases the wage rate by approximately 9.9%. A 95% interval estimate for the value of an additional year of education is 8.9% to 10.89%.

4.5.1 Prediction in the Log-Linear Model

You may have noticed that when reporting regression results in this section, we did not include an R^2 value. In a log-linear regression, the R^2 value automatically reported by statistical software is the percentage of the variation in $\ln(y)$ explained by the model. However, our objective is to explain the variations in y, not $\ln(y)$. Furthermore, the fitted regression line predicts $\widehat{\ln(y)} = b_1 + b_2 x$, whereas we want to predict y. The problems of obtaining a useful measure of goodness-of-fit and prediction are connected, as we discussed in Section 4.2.2.

How shall we obtain the predicted value of y? A first inclination might be to take the antilog of $\widehat{\ln(y)} = b_1 + b_2 x$. The exponential function is the antilogarithm for the natural logarithm, so that a natural choice for prediction is

$$\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp\left(b_1 + b_2 x\right)$$

In the log-linear model, this is not necessarily the best we can do. Using properties of the log-normal distribution it can be shown (see Appendix B.3.9) that an alternative predictor is

$$\hat{y}_c = \widehat{E(y)} = \exp\left(b_1 + b_2 x + \hat{\sigma}^2/2\right) = \hat{y}_n e^{\hat{\sigma}^2/2}$$

If the sample size is large, the "corrected" predictor $\hat{y}_c$ is, on average, closer to the actual value of y and should be used. In small samples (less than 30), the "natural" predictor may actually be a better choice. The reason for this incongruous result is that the estimated value of the error variance $\hat{\sigma}^2$ adds a certain amount of "noise" when using $\hat{y}_c$, leading it to have increased variability relative to $\hat{y}_n$ that can outweigh the benefit of the correction in small samples.

EXAMPLE 4.11 | Prediction in a Log-Linear Model

The effect of the correction can be illustrated using the wage equation. What would we predict the wage to be for a worker with 12 years of education? The predicted value of ln(*WAGE*) is

$$\widehat{\ln(WAGE)} = 1.5968 + 0.0988 \times EDUC$$
$$= 1.5968 + 0.0988 \times 12 = 2.7819$$

Then the value of the natural predictor is $\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(2.7819) = 16.1493$. The value of the corrected predictor, using $\hat{\sigma}^2 = 0.2349$ from the regression output, is

$$\hat{y}_c = \widehat{E(y)} = \hat{y}_n e^{\hat{\sigma}^2/2} = 16.1493 \times 1.1246 = 18.1622$$

We predict that the wage for a worker with 12 years of education will be \$16.15 per hour if we use the natural predictor and \$18.16 if we use the corrected predictor. In this case, the sample is large ($N = 1200$), so we would use the corrected predictor. Among the 1200 workers, there are 307 with 12 years of education. Their average wage is \$17.31, so the corrected predictor is consistent with the sample of data.

How does the correction affect our prediction? Recall that $\hat{\sigma}^2$ must be greater than zero and $e^0 = 1$. Thus, the effect of the correction is always to increase the value of the prediction because $e^{\hat{\sigma}^2/2}$ is always greater than one. The natural predictor tends to systematically underpredict the value of y in a log-linear model, and the correction offsets the downward bias in large samples. The "natural" and "corrected" predictions are shown in Figure 4.13.

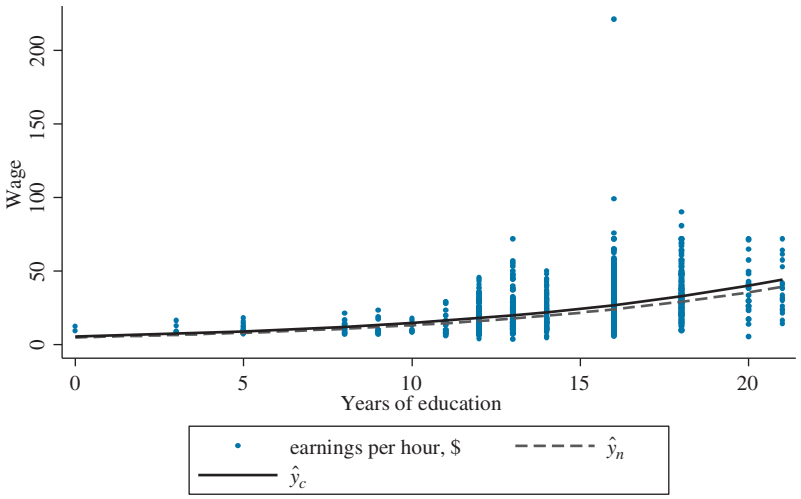


FIGURE 4.13 **The natural and corrected predictors of wage.**

4.5.2 A Generalized R^2 Measure

It is a general rule that the squared simple correlation between y and its fitted value $\hat{y}$, where $\hat{y}$ is the "best" prediction one can obtain, is a valid measure of goodness-of-fit that we can use as an R^2 in many contexts. As we have seen, what we may consider the "best" predictor can change depending on the model under consideration. That is, a general goodness-of-fit measure, or general R^2, is

$$R_g^2 = \left[\text{corr}\left(y, \hat{y}\right)\right]^2 = r_{y\hat{y}}^2$$

In the wage equation $R_g^2 = \left[\text{corr}\left(y, \hat{y}_c\right)\right]^2 = 0.4647^2 = 0.2159$, as compared to the reported $R^2 = 0.2577$ from the regression of ln(*WAGE*) on *EDUC*. (In this case since the corrected and natural predictors differ only by a constant factor, the correlation is the same for both.) These R^2 values are small, but we repeat our earlier message: R^2 values tend to be small with microeconomic, cross-sectional data because the variations in individual behavior are difficult to fully explain.

4.5.3 Prediction Intervals in the Log-Linear Model

We have a corrected predictor $\hat{y}_c$ for y in the log-linear model. It is the "point" predictor, or point forecast, that is relevant if we seek the single number that is our best prediction of y.

If we prefer a prediction or forecast interval for y, then we must rely on the natural predictor $\hat{y}_n$.[1] Specifically, we follow the procedure outlined in Section 4.1 and then take antilogs. That is, compute $\widehat{\ln(y)} = b_1 + b_2 x$ and then $\widehat{\ln(y)} \pm t_c \text{se}(f)$, where the critical value t_c is the $100(1-\alpha/2)$-percentile from the t-distribution and $\text{se}(f)$ is given in (4.5). Then, a $100(1-\alpha)\%$ prediction interval for y is

$$\left[\exp\left(\widehat{\ln(y)} - t_c \text{se}(f)\right), \exp\left(\widehat{\ln(y)} + t_c \text{se}(f)\right)\right]$$

EXAMPLE 4.12 | Prediction Intervals for a Log-Linear Model

For the wage data, a 95% prediction interval for the wage of a worker with 12 years of education is

$$\begin{aligned}&\left[\exp(2.7819 - 1.96 \times 0.4850), \exp(2.7819 + 1.96 \times 0.4850)\right] \\ &= [6.2358, 41.8233]\end{aligned}$$

The interval prediction is \$6.24–\$41.82, which is so wide that it is basically useless. What does this tell us? Nothing we did not already know. Our model is not an accurate predictor of individual behavior in this case. In later chapters, we will see if we can improve this model by adding additional explanatory variables, such as experience, that should be relevant. The prediction interval is shown in Figure 4.14.

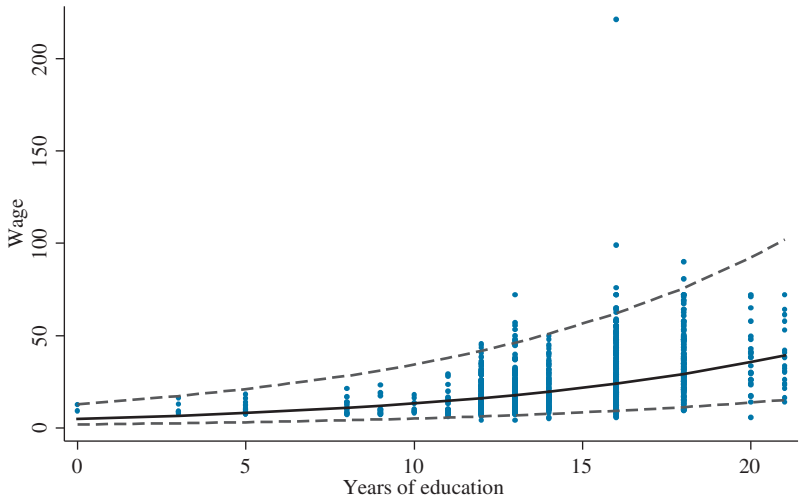


FIGURE 4.14 **The 95% prediction interval for wage.**

4.6 Log-Log Models

The log-log function, $\ln(y) = \beta_1 + \beta_2 \ln(x)$, is widely used to describe demand equations and production functions. The name "log-log" comes from the fact that the logarithm appears on both sides of the equation. In order to use this model, all values of y and x must be positive. Using the

[1] See Appendix 4A. The corrected predictor includes the estimated error variance, making the t-distribution no longer relevant in (4A.1).

properties of logarithms, we can see how to interpret the parameter of a log-log model. Consider an increase in x from x_0 to x_1. The change in the log-log model is from $\ln(y_0) = \beta_1 + \beta_2 \ln(x_0)$ to $\ln(y_1) = \beta_1 + \beta_2 \ln(x_1)$. Subtracting the first equation from the second gives $\ln(y_1) - \ln(y_0) = \beta_2[\ln(x_1) - \ln(x_0)]$. Multiply by 100, and use the approximation introduced in Appendix A, equation (A.3) to obtain $100[\ln(y_1) - \ln(y_0)] \cong \%\Delta y$ and $100[\ln(x_1) - \ln(x_0)] \cong \%\Delta x$, so that $\%\Delta y = \beta_2 \%\Delta x$, or $\beta_2 = \%\Delta y / \%\Delta x = \varepsilon_{yx}$. That is, in the log-log model, the parameter β_2 is the elasticity of y with respect to a change in x, and it is constant over the entire curve.

A useful way to think about the log-log model comes from a closer inspection of its slope. The slope of the log-log model changes at every point, and it is given by $dy/dx = \beta_2(y/x)$. Rearrange this so that $\beta_2 = (dy/y)/(dx/x)$. Thus, the slope of the log-log function exhibits constant *relative* change, whereas the linear function displays constant absolute change. The log-log function is a transformation of the equation $y = Ax^{\beta_2}$, with $\beta_1 = \ln(A)$. The various shape possibilities for log-log models are depicted in Figure 4.5(c), for $\beta_2 > 0$, and Figure 4.5(d), for $\beta_2 < 0$.

If $\beta_2 > 0$, then y is an increasing function of x. If $\beta_2 > 1$, then the function increases at an increasing rate. That is, as x increases the slope increases as well. If $0 < \beta_2 < 1$, then the function is increasing, but at a decreasing rate; as x increases, the slope decreases.

If $\beta_2 < 0$, then there is an inverse relationship between y and x. If, for example, $\beta_2 = -1$, then $y = Ax^{-1}$ or $xy = A$. This curve has "unit" elasticity. If we let y = quantity demanded and x = price, then A = total revenue from sales. For every point on the curve $xy = A$, the area under the curve A (total revenue for the demand curve) is constant. By definition, unit elasticity implies that a 1% increase in x (price, for example) is associated with a 1% decrease in y (quantity demanded), so that the product xy (price times quantity) remains constant.

EXAMPLE 4.13 | A Log-Log Poultry Demand Equation

The log-log functional form is frequently used for demand equations. Consider, for example, the demand for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the data file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006).[2] The scatter plot of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken is shown in Figure 4.15 for 52 annual observations, 1950–2001. It shows the characteristic hyperbolic shape that was displayed in Figure 4.5(d).

The estimated log-log model is

$$\widehat{\ln(Q)} = 3.717 - 1.121 \times \ln(P) \quad R_g^2 = 0.8817 \qquad (4.15)$$
$$\text{(se)} \quad (0.022) \quad (0.049)$$

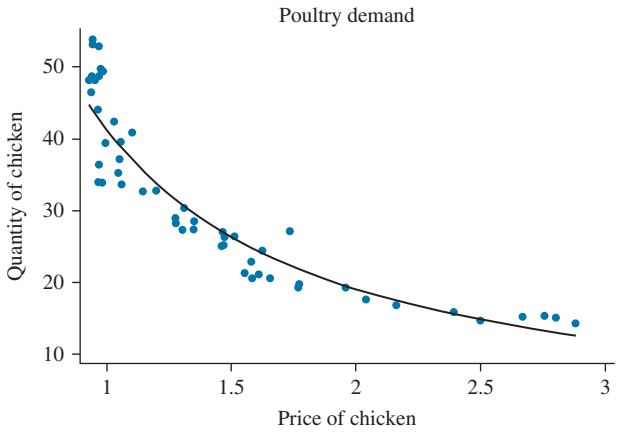


FIGURE 4.15 **Quantity and price of chicken.**

[2] "Simultaneous Equation Econometrics: The Missing Example," *Economic Inquiry*, 44(2), 374–384.

We estimate that the price elasticity of demand is 1.121: a 1% increase in real price is estimated to reduce quantity consumed by 1.121%.

The fitted line shown in Figure 4.15 is the "corrected" predictor discussed in Section 4.5.3. The corrected predictor $\hat{Q}_c$ is the natural predictor $\hat{Q}_n$ adjusted by the factor $\exp\left(\hat{\sigma}^2/2\right)$. That is, using the estimated error variance $\hat{\sigma}^2 = 0.0139$, the predictor is

$$\hat{Q}_c = \hat{Q}_n e^{\hat{\sigma}^2/2} = \exp\left(\widehat{\ln(Q)}\right) e^{\hat{\sigma}^2/2}$$
$$= \exp\left(3.717 - 1.121 \times \ln(P)\right) e^{0.0139/2}$$

The goodness-of-fit statistic $R_g^2 = 0.8817$ is the generalized R^2 discussed in Section 4.5.4. It is the squared correlation between the predictor $\hat{Q}_c$ and the observations Q

$$R_g^2 = \left[\text{corr}\left(Q, \hat{Q}_c\right)\right]^2 = [0.939]^2 = 0.8817$$

4.7 Exercises

4.7.1 Problems

4.1 Answer each of the following:

a. Suppose that a simple regression has quantities $N = 20$, $\sum y_i^2 = 7825.94$, $\bar{y} = 19.21$, and $SSR = 375.47$, find R^2.

b. Suppose that a simple regression has quantities $R^2 = 0.7911$, $SST = 725.94$, and $N = 20$, find $\hat{\sigma}^2$.

c. Suppose that a simple regression has quantities $\sum\left(y_i - \bar{y}\right)^2 = 631.63$ and $\sum \hat{e}_i^2 = 182.85$, find R^2.

4.2 Consider the following estimated regression equation (standard errors in parentheses):

$$\hat{y} = 64.29 + 0.99x \quad R^2 = 0.379$$
$$\text{(se)} \quad (2.42) \quad (0.18)$$

Rewrite the estimated equation, including coefficients, standard errors, and R^2, that would result if

a. All values of x were divided by 10 before estimation.

b. All values of y were divided by 10 before estimation.

c. All values of y and x were divided by 10 before estimation.

4.3 We have five observations on x and y. They are $x_i = 3, 2, 1, -1, 0$ with corresponding y values $y_i = 4, 2, 3, 1, 0$. The fitted least squares line is $\hat{y}_i = 1.2 + 0.8x_i$, the sum of squared least squares residuals is $\sum_{i=1}^5 \hat{e}_i^2 = 3.6$, $\sum_{i=1}^5 \left(x_i - \bar{x}\right)^2 = 10$, and $\sum_{i=1}^5 \left(y_i - \bar{y}\right)^2 = 10$. Carry out this exercise with a hand calculator. Compute

a. the predicted value of y for $x_0 = 4$.

b. the se(f) corresponding to part (a).

c. a 95% prediction interval for y given $x_0 = 4$.

d. a 99% prediction interval for y given $x_0 = 4$.

e. a 95% prediction interval for y given $x = \bar{x}$. Compare the width of this interval to the one computed in part (c).

4.4 The general manager of a large engineering firm wants to know whether the experience of technical artists influences their work quality. A random sample of 50 artists is selected. Using years of work experience (*EXPER*) and a performance rating (*RATING*, on a 100-point scale), two models are estimated by least squares. The estimates and standard errors are as follows:

Model 1:

$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$\text{(se)} \qquad (2.422) \quad (0.183)$$

Model 2:

$$\widehat{RATING} = 39.464 + 15.312\ln(EXPER) \quad N = 46 \quad R^2 = 0.6414$$
$$\text{(se)} \qquad (4.198) \quad (1.727)$$

a. Sketch the fitted values from Model 1 for $EXPER = 0$ to 30 years.
b. Sketch the fitted values from Model 2 against $EXPER = 1$ to 30 years. Explain why the four artists with no experience are not used in the estimation of Model 2.
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
d. Using Model 2, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

4.5 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$. *WAGE* is hourly wage rate in U.S. 2013 dollars. *EDUC* is years of education attainment, or schooling. The model is estimated using individuals from an urban area.

$$\widehat{WAGE} = -10.76 + 2.461965EDUC, \quad N = 986$$
$$\text{(se)} \qquad (2.27)\ (0.16)$$

a. The sample standard deviation of *WAGE* is 15.96 and the sum of squared residuals from the regression above is 199,705.37. Compute R^2.
b. Using the answer to (a), what is the correlation between *WAGE* and *EDUC*? [*Hint*: What is the correlation between *WAGE* and the fitted value $\widehat{WAGE}$?]
c. The sample mean and variance of *EDUC* are 14.315 and 8.555, respectively. Calculate the leverage of observations with $EDUC = 5$, 16, and 21. Should any of the values be considered large?
d. Omitting the ninth observation, a person with 21 years of education and wage rate \$30.76, and reestimating the model we find $\hat{\sigma} = 14.25$ and an estimated slope of 2.470095. Calculate DFBETAS for this observation. Should it be considered large?
e. For the ninth observation, used in part (d), DFFITS $= -0.0571607$. Is this value large? The leverage value for this observation was found in part (c). How much does the fitted value for this observation change when this observation is deleted from the sample?
f. For the ninth observation, used in parts (d) and (e), the least squares residual is -10.18368. Calculate the studentized residual. Should it be considered large?

4.6 We have five observations on x and y. They are $x_i = 3, 2, 1, -1, 0$ with corresponding y values $y_i = 4, 2, 3, 1, 0$. The fitted least squares line is $\hat{y}_i = 1.2 + 0.8x_i$, the sum of squared least squares residuals is $\sum_{i=1}^{5} \hat{e}_i^2 = 3.6$ and $\sum_{i=1}^{5} (y_i - \bar{y})^2 = 10$. Carry out this exercise with a hand calculator.

a. Calculate the fitted values $\hat{y}_i$ and their sample mean $\bar{\hat{y}}$. Compare this value to the sample mean of the y values.
b. Calculate $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2$ and $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{5} (y_i - \bar{y})^2$.
c. The least squares residuals are $\hat{e}_i = 0.4, -0.8, 1, 0.6$, and -1.2. Calculate $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})\hat{e}_i$.
d. Calculate $1 - \sum_{i=1}^{5} \hat{e}_i^2 / \sum_{i=1}^{5} (y_i - \bar{y})^2$ and compare it to the results in part (b).
e. Show, algebraically, that $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^{5} \hat{y}_i y_i - N\bar{y}^2$. Calculate this value.
f. Using $\sum_{i=1}^{5} (x_i - \bar{x})^2 = 10$, and previous results, calculate

$$r = \left[\sum_{i=1}^{5} (\hat{y}_i - \bar{y})(y_i - \bar{y})\right] \Big/ \left[\sqrt{\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^{5} (y_i - \bar{y})^2}\right]$$

What statistic is r? Calculate r^2 and compare this value to the values in parts (d) and (b).

4.7 We have data on 2323 randomly selected households consisting of three persons in 2013. Let *ENTERT* denote the monthly entertainment expenditure (\$) per person per month and let *INCOME* (\$100) be monthly household income. Consider the regression model

$$ENTERT_i = \beta_1 + \beta_2 INCOME_i + e_i, \quad i = 1, \ldots, 2323$$

Assume that assumptions SR1–SR6 hold. The OLS estimated equation is $\widehat{ENTERT}_i = 9.820 + 0.503INCOME_i$. The standard error of the slope coefficient estimator is $\text{se}(b_2) = 0.029$, the standard

error of the intercept estimator is $\text{se}(b_1) = 2.419$, and the estimated covariance between the least squares estimators b_1 and b_2 is -0.062. From the summary statistics, we find

$$\sum_{i=1}^{2323}\left(ENTERT_i - \overline{ENTERT}\right)^2 = 8691035,\ \sum_{i=1}^{2323}\left(INCOME_i - \overline{INCOME}\right)^2 = 3876440$$
$$\overline{ENTERT} = 45.93,\ \overline{INCOME} = 71.84$$

a. From the estimated regression, the sum of squared least squares residuals is 7711432. How well does the regression model fit the data? How much of the household variation in entertainment expenses have we explained using this regression model? Explain your answer.

b. The Jones household has income of \$10,000 per month. Predict their per person household expenditure on entertainment.

c. Calculate a 95% prediction interval for the Jones household's per person expenditure on entertainment. Show your work.

d. Calculate a 95% prediction interval for the Jones household's total household expenditure on entertainment. Show your work.

4.8 Consider a log-linear regression for the weekly sales (number of cans) of a national brand of canned tuna ($SAL1$ = target brand sales) as a function of the ratio of its price to the price of a competitor, $RPRICE3 = 100$(price of target brand ÷ price competitive brand #3), $\ln(SAL1) = \gamma_1 + \gamma_2 RPRICE3 + e$. Using $N = 52$ weekly observations, the OLS estimated equation is

$$\widehat{\ln(SAL1)} = 11.481 - 0.031 RPRICE3$$
$$\text{(se)} \qquad (0.535) \quad (0.00529)$$

a. The sample mean of $RPRICE3$ is 99.66, its median is 100, its minimum value is 70.11, and its maximum value is 154.24. What do these summary statistics tell us about the prices of the target brand relative to the prices of its competitor?

b. Interpret the coefficient of $RPRICE3$. Does its sign make economic sense?

c. Using the "natural" predictor, predict the weekly sales of the target brand if $RPRICE3$ takes its sample mean value. What is the predicted sales if $RPRICE3$ equals 140?

d. The estimated value of the error variance from the regression above is $\hat{\sigma}^2 = 0.405$ and $\sum_{i=1}^{52}\left(RPRICE3_i - \overline{RPRICE3}\right)^2 = 14757.57$. Construct a 90% prediction interval for the weekly sales of the target brand if $RPRICE3$ takes its sample mean value. What is the 90% prediction interval for sales if $RPRICE3$ equals 140? Is one interval wider? Explain why this happens.

e. The fitted value of $\ln(SAL1)$ is $\widehat{\ln(SAL1)}$. The correlation between $\ln(SAL1)$ and $\widehat{\ln(SAL1)}$ is 0.6324, the correlation between $\widehat{\ln(SAL1)}$ and $SAL1$ is 0.5596, and the correlation between $\exp\left[\widehat{\ln(SAL1)}\right]$ and $SAL1$ is 0.6561. Calculate the R^2 that would normally be shown with the fitted regression output above. What is its interpretation? Calculate the "generalized-R^2." What is its interpretation?

4.9 Consider the weekly sales (number of cans) of a national brand of canned tuna ($SAL1$ = target brand sales) as a function of the ratio of its price to the price of a competitor, $RPRICE3 = 100$(price of target brand ÷ price competitive brand #3). Using $N = 52$ weekly observations, and for this exercise scaling $SAL1/1000$ so that we have sales measured as thousands of cans per week, we obtain the following least squares estimated equations, the first being a linear specification, the second a log-linear specification, and the third a log-log specification.

$$\widehat{SAL1} = 29.6126 - 0.2297 RPRICE3 \qquad \widehat{\ln(SAL1)} = 4.5733 - 0.0305 RPRICE3$$
$$\text{(se)} \quad (4.86) \quad (4.81) \qquad\qquad \text{(se)} \quad (0.54) \quad (0.0053)$$

$$\widehat{\ln(SAL1)} = 16.6806 - 3.3020 \ln(RPRICE3)$$
$$\text{(se)} \qquad (2.413) \quad (0.53)$$

a. For the linear specification, the sum of squared residuals is 1674.92, the estimated skewness and kurtosis of the residuals are 1.49 and 5.27, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance. Specify the distribution of the test statistic if the null hypothesis of normality is true and the rejection region.

b. For the log-linear specification, the estimated skewness and kurtosis of the residuals are 0.41 and 2.54, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance.

c. For the log-log specification, the estimated skewness and kurtosis of the residuals are 0.32 and 2.97, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance.

d. For the log-linear and log-log specifications, define a residual as $SAL1 - \exp\left(\widehat{\ln(SAL1)}\right)$. For the two models, the sum of the squared residuals as defined are 1754.77 for the log-linear model and 1603.14 for the log-log model. Based on these values, and comparing them to the sum of squared residuals from the linear specification, which model seems to fit the data best?

e. Table 4.2 reports correlations between the regression model variables and predictions from the linear relationship (*YHAT*), predictions from the log-linear relationship $\left(YHATL = \exp\left[\widehat{\ln(SAL1)}\right]\right)$, and predictions from the log-log model $\left(YHATLL = \exp\left[\widehat{\ln(SAL1)}\right]\right)$.

- **i.** Why is the correlation between *SAL*1 and *RPRICE*3 the same as the correlation between *YHAT* and *SAL*1 (except for the sign)?
- **ii.** What is the R^2 from the linear relationship model?
- **iii.** Why is the correlation between *YHAT* and *RPRICE*3 a perfect—1.0?
- **iv.** What is the generalized-R^2 for the log-linear model?
- **v.** What is the generalized-R^2 for the log-log model?

f. Given the information provided in parts (a)–(e) which model would you select as having the best fit to the data?

TABLE 4.2 Correlations for Exercise 4.9

	*RPRICE*3	*SAL*1	*YHAT*	*YHATL*	*YHATLL*
*RPRICE*3	1.0000				
*SAL*1	−0.5596	1.0000			
YHAT	−1.0000	0.5596	1.0000		
YHATL	−0.9368	0.6561	0.9368	1.0000	
YHATLL	−0.8936	0.6754	0.8936	0.9927	1.0000

4.10 Using data on 76 countries, we estimate a relationship between the growth rate in prices, *INFLAT*, and the rate of growth in the money supply, *MONEY*. The least squares estimates of the model are as follows:

$$\begin{aligned} INFLAT &= -5.57 + 1.05MONEY \quad R^2 = 0.9917 \\ (\text{se}) \quad &\quad (0.70)\ (0.11) \end{aligned}$$

The data summary statistics are as follows:

	Mean	Median	Std. Dev.	Min	Max
INFLAT	25.35	8.65	58.95	−0.6	374.3
MONEY	29.59	16.35	56.17	2.5	356.7

Table 4.3 contains the data and some diagnostics for several observations.

a. Determine observations for which *LEVERAGE* is large. What is your rule?

b. Determine observations for which *EHATSTU* (the studentized residual) is large. What is your rule?

c. Determine observations for which *DFBETAS* is large. What is your rule?

d. Determine observations for which *DFFITS* is large. What is your rule?

e. Sketch the fitted relationship. On the graph locate the point of the means, and medians, and the nine data points in Table 4.3. Which observations are remote, relative to the center of the data, the point of the means and medians?

TABLE 4.3 Diagnostics for Selected Observations for Exercise 4.10

ID	INFLAT	MONEY	LEVERAGE	EHATSTU	DFBETAS	DFFITS
1	374.3	356.7	0.4654	1.8151	1.6694	1.6935
2	6.1	11.5	0.0145	−0.0644	0.0024	−0.0078
3	3.6	7.3	0.0153	0.2847	−0.0131	0.0354
4	187.1	207.1	0.1463	−5.6539	−2.2331	−2.3408
5	12.3	25.2	0.0132	−1.5888	0.0144	−0.1840
6	4.0	3.1	0.0161	1.1807	−0.0648	0.1512
7	316.1	296.6	0.3145	2.7161	1.8007	1.8396
8	13.6	17.4	0.0138	0.1819	−0.0046	0.0215
9	16.4	18.5	0.0137	0.4872	−0.0112	0.0574

4.11 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$ where *WAGE* is hourly wage rate in U.S. 2013 dollars, *EDUC* is years of education attainment. The model is estimated twice, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \quad \underset{\text{(se)}}{\widehat{WAGE} = -10.76 + 2.46EDUC,} \quad \underset{(2.27)\ (0.16)}{N = 986}$$

$$\text{Rural} \quad \underset{\text{(se)}}{\widehat{WAGE} = -4.88 + 1.80EDUC,} \quad \underset{(3.29)\ (0.24)}{N = 214}$$

a. For the rural regression, compute a 95% prediction interval for *WAGE* if $EDUC = 16$, and the standard error of the forecast is 9.24. The standard error of the regression is $\hat{\sigma} = 9.20$ for the rural data.

b. For the urban data, the sum of squared deviations of *EDUC* about its sample mean is 8435.46 and the standard error of the regression is $\hat{\sigma} = 14.25$. The sample mean wage in the urban area is \$24.49. Calculate the 95% prediction interval for *WAGE* if $EDUC = 16$. Is the interval wider or narrower than the prediction interval for the rural data? Do you find this plausible? Explain.

4.12 Consider the share of total household expenditure (*TOTEXP*) devoted to expenditure on food (*FOOD*). Specify the log-linear relationship $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$.

a. Show that the elasticity of expenditure on food with respect to total expenditure is

$$\varepsilon = \frac{dFOOD}{dTOTEXP} \times \frac{TOTEXP}{FOOD} = \frac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

[*Hint*: Solve the log-linear relationship as $FOOD = \left[\beta_1 + \beta_2 \ln(TOTEXP)\right] TOTEXP$ and differentiate to obtain $dFOOD/dTOTEXP$. Then multiply by $TOTEXP/FOOD$ and simplify.]

b. The least squares estimates of the regression model $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP) + e$, using 925 observations from London, are as follows:

$$\underset{(t)}{\widehat{\frac{FOOD}{TOTEXP}} = 0.953 - 0.129 \ln(TOTEXP)} \quad \underset{(26.10)\,(-16.16)}{R^2 = 0.2206,} \quad \hat{\sigma} = 0.0896$$

Interpret the estimated coefficient of $\ln(TOTEXP)$. What happens to the share of food expenditure in the budget as total household expenditures increase?

c. Calculate the elasticity in part (a) at the 5th percentile, and the 75th percentile of total expenditure. Is this a constant elasticity function? The 5th percentile is 500 UK pounds, and the 75th percentile is 1200 UK pounds.

d. The residuals from the model in (b) have skewness 0.0232 and kurtosis 3.4042. Carry out the Jarque–Bera test at the 1% level of significance. What are the null and alternative hypotheses for this test?

e. In $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$, take the logarithm of the left-hand side and simplify the result to obtain $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP)$. How are the parameters in this model related to the budget share relation?

f. The least squares estimates of $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$ are as follows:

$$\widehat{\ln(FOOD)} = 0.732 + 0.608 \ln(TOTEXP) \quad R^2 = 0.4019 \quad \hat{\sigma} = 0.2729$$
$$(t) \qquad (6.58) \ (24.91)$$

Interpret the estimated coefficient of $\ln(TOTEXP)$. Calculate the elasticity in this model at the 5th percentile and the 75th percentile of total expenditure. Is this a constant elasticity function?

g. The residuals from the log-log model in (e) show skewness $= -0.887$ and kurtosis $= 5.023$. Carry out the Jarque–Bera test at the 5% level of significance.

h. In addition to the information in the previous parts, we multiply the fitted value in part (b) by *TOTEXP* to obtain a prediction for expenditure on food. The correlation between this value and actual food expenditure is 0.641. Using the model in part (e) we obtain $\exp\left[\widehat{\ln(FOOD)}\right]$. The correlation between this value and actual expenditure on food is 0.640. What if any information is provided by these correlations? Which model would you select for reporting, if you had to choose only one? Explain your choice.

4.13 The linear regression model is $y = \beta_1 + \beta_2 x + e$. Let $\bar{y}$ be the sample mean of the y-values and $\bar{x}$ the average of the x-values. Create variables $\tilde{y} = y - \bar{y}$ and $\tilde{x} = x - \bar{x}$. Let $\tilde{y} = \alpha \tilde{x} + e$.

a. Show, algebraically, that the least squares estimator of α is identical to the least square estimator of β_2. [*Hint*: See Exercise 2.4.]

b. Show, algebraically, that the least squares residuals from $\tilde{y} = \alpha \tilde{x} + e$ are the same as the least squares residuals from the original linear model $y = \beta_1 + \beta_2 x + e$.

4.14 Using data on 5766 primary school children, we estimate two models relating their performance on a math test (*MATHSCORE*) to their teacher's years of experience (*TCHEXPER*).

Linear relationship

$$\widehat{MATHSCORE} = 478.15 + 0.81 TCHEXPER \quad R^2 = 0.0095 \quad \hat{\sigma} = 47.51$$
$$(se) \qquad (1.19) \ (0.11)$$

Linear-log relationship

$$\widehat{MATHSCORE} = 474.25 + 5.63 \ln(TCHEXPER) \quad R^2 = 0.0081 \quad \hat{\sigma} = 47.57$$
$$(se) \qquad (1.84) \ (0.84)$$

a. Using the linear fitted relationship, how many years of additional teaching experience is required to increase the expected math score by 10 points? Explain your calculation.

b. Does the linear fitted relationship imply that at some point there are diminishing returns to additional years of teaching experience? Explain.

c. Using the fitted linear-log model, is the graph of *MATHSCORE* against *TCHEXPER* increasing at a constant rate, at an increasing rate, or at a decreasing rate? Explain. How does this compare to the fitted linear relationship?

d. Using the linear-log fitted relationship, if a teacher has only one year of experience, how many years of extra teaching experience is required to increase the expected math score by 10 points? Explain your calculation.

e. 252 of the teachers had no teaching experience. What effect does this have on the estimation of the two models?

f. These models have such a low R^2 that there is no statistically significant relationship between expected math score and years of teaching experience. True or False? Explain your answer.

4.15 Consider a **log-reciprocal model** that relates the logarithm of the dependent variable to the reciprocal of the explanatory variable, $\ln(y) = \beta_1 + \beta_2(1/x)$. [Note: An illustration of this model is given in Exercise 4.17].

a. For what values of y is this model defined? Are there any values of x that cause problems?

b. Write the model in exponential form as $y = \exp\left[\beta_1 + \beta_2(1/x)\right]$. Show that the slope of this relationship is $dy/dx = \exp\left[\beta_1 + (\beta_2/x)\right] \times (-\beta_2/x^2)$. What sign must β_2 have for y and x to have a positive relationship, assuming that $x > 0$?

c. Suppose that $x > 0$ but it converges toward zero from above. What value does y converge to? What does y converge to as x approaches infinity?

d. Suppose $\beta_1 = 0$ and $\beta_2 = -4$. Evaluate the slope at the x-values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0. As x increases, is the slope of the relationship increasing or decreasing, or both?

e. Show that the second derivative of the function is

$$\frac{d^2y}{dx^2} = \left(\frac{\beta_2^2}{x^4} + \frac{2\beta_2}{x^3}\right)\exp\left[\beta_1 + (\beta_2/x)\right]$$

Assuming $\beta_2 < 0$ and $x > 0$, set the equation to zero, and show that the x-value that makes the second derivative zero is $-\beta_2/2$. Does this result agree with your calculations in part (d)? [*Hint*: $\exp\left[\beta_1 + (\beta_2/x)\right] > 0$. You have solved for what is called an *inflection point*.]

4.7.2 Computer Exercises

4.16 In Section 4.6, we considered the demand for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*.

a. Using the 52 annual observations, 1950–2001, estimate the **reciprocal model** $Q = \alpha_1 + \alpha_2(1/P) + e$. Plot the fitted value of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

b. Using the estimated relation in part (a), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31, and quantity is taken to be the corresponding value on the fitted curve.

[*Hint*: The derivative (slope) of the reciprocal model $y = a + b(1/x)$ is $dy/dx = -b(1/x^2)$.] Compare this estimated elasticity to the estimate found in Section 4.6 where the log-log functional form was used.

c. Estimate the poultry consumption using the linear-log functional form $Q = \gamma_1 + \gamma_2 \ln(P) + e$. Plot the fitted values of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

d. Using the estimated relation in part (c), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the log-log model and from the reciprocal model in part (b).

e. Estimate the poultry consumption using a log-linear model $\ln(Q) = \phi_1 + \phi_2 P + e$. Plot the fitted values of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

f. Using the estimated relation in part (e), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the previous models.

g. Evaluate the suitability of the alternative models for fitting the poultry consumption data, including the log-log model. Which of them would you select as best, and why?

4.17 McCarthy and Ryan (1976) considered a model of television ownership in the United Kingdom and Ireland using data from 1955 to 1973. Use the data file *tvdata* for this exercise.

a. For the United Kingdom, plot the rate of television ownership (*RATE_UK*) against per capita consumer expenditures (*SPEND_UK*). Which models in Figure 4.5 are candidates to fit the data?

b. Estimate the linear-log model $RATE\_UK = \beta_1 + \beta_2 \ln(SPEND\_UK) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data?

c. What is the interpretation of the intercept in the linear-log model? Specifically, for the model in (b), for what value of *SPEND_UK* is the expected value $E(RATE\_UK|SPEND\_UK) = \beta_1$?

d. Estimate the linear-log model $RATE\_UK = \beta_1 + \beta_2 \ln(SPEND\_UK - 280) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data? How has the adjustment (−280) changed the fitted relationship? [Note: You might well wonder how the value 280 was determined. It was estimated using a procedure called *nonlinear least squares*. You will be introduced to this technique later in this book.]

e. A competing model is the log-reciprocal model, described in Exercise 4.15. Estimate the log-reciprocal model $\ln(RATE\_UK) = \alpha_1 + \alpha_2(1/SPEND\_UK) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data?

f. Explain the failure of the model in (e) by referring to Exercise 4.15(c).

g. Estimate the log-reciprocal model $\ln(RATE\_UK) = \alpha_1 + \alpha_2(1/[SPEND\_UK - 280]) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data? How has this modification corrected the problem identified in part (f)?

h. Repeat the above exercises for Ireland, with correcting factor 240 instead of 280.

4.18 Do larger universities have lower cost per student or a higher cost per student? Use the data on 141 public universities in the data file *pubcoll* for 2010 and 2011. A university is many things and here we only focus on the effect of undergraduate full-time student enrollment (*FTESTU*) on average total cost per student (*ACA*). Consider the regression model $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + e_{it}$ where the subscripts i and t denote the university and the time period, respectively. Here, e_{it} is the usual random error term.

a. Estimate the model above using 2010 data only, again using 2011 data only, and again using both years of data together. What is the estimated effect of increasing enrollment on average cost per student? Base your answer on both point and 95% interval estimates.

b. There are certainly many other factors affecting average cost per student. Some of them can be characterized as the university "identity" or "image." Let us denote these largely unobservable individual attributes as u_i. If we could add this feature to the model, it would be $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + (\theta u_i + e_{it})$. We place it in parentheses with e_{it} because it is another unobservable random error, but it is different because the character or identify of a university does not change from one year to the next. Do you suppose that our usual exogeneity assumptions hold in light of this new class of omitted variables? Might some unobservable characteristics of a university be correlated with student enrollment? Give some examples.

c. With our two years of data, we can take "first differences," by subtracting the model in 2010 from the model in 2011, $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta e_i$, where

$$\Delta ACA_i = ACA_{i,2011} - ACA_{i,2010}$$
$$\Delta FTESTU_i = FTESTU_{i,2011} - FTESTU_{i,2010}$$
$$\Delta e_i = e_{i,2011} - e_{i,2010}$$

Explain why the intercept and θu_i drop from the model. Explain how the exogeneity assumptions might now hold.

d. Estimate $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta e_i$ and also $\Delta ACA_i = \delta + \beta_2 \Delta FTESTU_i + \Delta e_i$. What now is the estimated effect of increasing enrollment on average cost per student? Base your answer on both point and 95% interval estimates. Does adding an intercept to the model make any fundamental difference in this case?

e. Estimate the model $\Delta\ln(ACA_i) = \alpha + \gamma \Delta\ln(FTESTU_i) + \Delta e_i$ where

$$\Delta\ln(ACA_i) = \ln(ACA_{i,2011}) - \ln(ACA_{i,2010})$$

and

$$\Delta\ln(FTESTU_i) = \ln(FTESTU_{i,2011}) - \ln(FTESTU_{i,2010})$$

Interpret the estimated coefficient of $\Delta\ln(FTESTU_i)$.

[*Hint*: See equation (A.3) in Appendix A.]

4.19 The data file *wa_wheat* contains wheat yield for several shires in Western Australia from 1950 to 1997.

a. If the variable *YIELD* is "average wheat yield" in tonnes per hectare what is the interpretation of $RYIELD = 1/YIELD$?

b. For Northampton and Mullewa shires, plot $RYIELD = 1/YIELD$ against $YEAR = 1949 + TIME$. Do you notice any anomalies in the plots? What years are most unusual? Using your favorite search engine discover what conditions may have affected wheat production in these shires during these years.

c. For Northampton and Mullewa shires, estimate the reciprocal model $RYIELD = \alpha_1 + \alpha_2 TIME + e$. Interpret the estimated coefficient. What does the sign tell us?

d. For the estimations in part (c), test the hypothesis that the coefficient of *TIME* is greater than or equal to zero against the alternative that it is negative, at the 5% level of significance.

e. For each of the estimations in part (c), calculate studentized residuals, and values for the diagnostics *LEVERAGE*, *DFBETAS*, and *DFFITS*. Identify the years in which these are "large" and include your threshold for what is large.

f. Discarding correct data is hardly ever a good idea, and we recommend that you not do it. Later in this book, you will discover other methods for addressing such problems—such as adding

additional explanatory variables—but for now experiment. For each shire, identify the most unusual observation. What grounds did you use for choosing?

g. Drop the most unusual observation for each shire and reestimate the model. How much do the results change? How do these changes relate to the diagnostics in part (e)?

4.20 In the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$, the corrected predictor $\hat{y}_c = \exp(b_1 + b_2 x) \times \exp\left(\hat{\sigma}^2/2\right)$ is argued to have a lower mean squared error than the "normal" predictor $\hat{y}_n = \exp(b_1 + b_2 x)$. The correction factor $\exp\left(\hat{\sigma}^2/2\right)$ depends on the regression errors having a normal distribution.

a. In exponential form, the log-linear model is $y = \exp(\beta_1 + \beta_2 x)\exp(e)$. Assuming that the explanatory variable x and the random error e are statistically independent, find $E(y)$.

b. Use the data file *cps5_small* for this exercise. [The data file *cps5* contains more observations and variables.] Estimate the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$ using the first 1000 observations. Based on this regression, calculate the correction factor $c = \exp\left(\hat{\sigma}^2/2\right)$. What is this value?

c. Obtain the 1000 least squares residuals $\hat{e}$ from the regression in (b). Calculate the correction factor $d = \sum_{i=1}^{1000} \exp(\hat{e}_i)/1000$. What is this value?

d. Using the estimates from part (b), obtain the predictions for observations 1001–1200, using $\hat{y}_n = \exp(b_1 + b_2 x)$, $\hat{y}_c = c\hat{y}_n$, and $\hat{y}_d = d\hat{y}_n$. Calculate the mean (average) squared forecast errors $MSE_n = \sum_{i=1001}^{1200} (\hat{y}_{ni} - y_i)^2/200$, $MSE_c = \sum_{i=1001}^{1200} (\hat{y}_{ci} - y_i)^2/200$, and $MSE_d = \sum_{i=1001}^{1200} (\hat{y}_{di} - y_i)^2/200$. Based on this criterion, which predictor is best?

4.21 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel.

a. Locate Malawi and its neighboring countries on a map. Find the exchange rate between US $1 and the Malawian Kwacha. What is the population size of Malawi? Which industry drives the Malawi economy?

b. Define the proportion of expenditure on food as $PFOOD = FOOD/TOTEXP$. Estimate the linear-log regression model $PFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e$ and report the estimation results. What happens to the share of total expenditure devoted to food as total expenditure rises. Construct a 95% interval estimate for β_2. Have we estimated this coefficient relatively precisely or not? Does the model fit the data well? Is there a problem?

c. The elasticity of expenditure on food with respect to total expenditure is

$$\varepsilon = \frac{dFOOD}{dTOTEXP} \times \frac{TOTEXP}{FOOD} = \frac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

This result is derived in Exercise 4.12. Calculate the elasticity at the 5th percentile and the 75th percentile of total expenditure. Is this a constant elasticity function? If your software permits, calculate a standard error for the elasticity.

d. Calculate the least squares residuals from the model in (b). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance. What are the null and alternative hypotheses for this test?

e. Take the logarithm of the left-hand side of $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$ and simplify the result, and add an error term, to obtain $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v$. Estimate this model. Interpret the estimated coefficient of $\ln(TOTEXP)$. What is the estimated elasticity of expenditure on food with respect to total expenditure?

f. Calculate the residuals from the model in (e). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance.

g. Estimate the linear-log model $FOOD = \gamma_1 + \gamma_2 \ln(TOTEXP) + u$. Discuss the estimation results. Calculate the elasticity of food expenditure with respect to total expenditure when food expenditure is at its 50th percentile and at its 75th percentile. Is this a constant elasticity function, or is elasticity increasing or decreasing?

h. Calculate the residuals from the model in (g). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance.

i. Calculate predicted values of expenditure on food from each model. Multiply the fitted value from the model in part (b) to obtain a prediction for expenditure on food. Using the model in part (e) obtain $\exp\left[\widehat{\ln(FOOD)}\right]$. For the model in part (g), obtain fitted values. Find the correlations between the actual value of *FOOD* and the three sets of predictions. What, if any, information is provided by these correlations? Which model would you select for reporting, if you had to choose only one? Explain your choice.

4.22 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel.

a. Define the proportion of expenditure on food consumed away from home as $PFOODAWAY = FOODAWAY/TOTEXP$. Construct a histogram for *PFOODAWAY* and its summary statistics. What percentage of the sample has a zero value for *PFOODAWAY*. What does that imply about their expenditures last month?

b. Create the variable $FOODAWAY = PFOODAWAY \times TOTEXP$. Construct a histogram for *FOODAWAY* and another histogram for *FOODAWAY* if $FOODAWAY > 0$. Compare the summary statistics for *TOTEXP* for households with $FOODAWAY > 0$ to those with $FOODAWAY = 0$. What differences do you observe?

c. Estimate the linear regression model $FOODAWAY = \beta_1 + \beta_2 TOTEXP + e$ twice, once for the full sample, and once using only households for whom $FOODAWAY > 0$. What differences in slope estimates do you observe? How would you explain these differences to an audience of noneconomists?

d. Calculate the fitted values from each of the estimated models in part (c) and plot the fitted values, and *FOODAWAY* values, versus *TOTEXP*. Think about how the least squares estimation procedure works to fit a line to data. Explain the relative difference in the two estimations based on this intuition.

4.23 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel. Consider the following models.

i. Budget share: $PTELEPHONE = \beta_1 + \beta_2 \ln(TOTEXP) + e$

ii. Expenditure: $\ln(PTELEPHONE \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

iii. Budget share: $PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e$

iv. Expenditure: $\ln(PCLOTHES \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

v. Budget share: $PFUEL = \beta_1 + \beta_2 \ln(TOTEXP) + e$

vi. Expenditure: $\ln(PFUEL \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

a. Estimate each of the models (i) to (vi). Interpret the estimated coefficients of $\ln(TOTEXP)$. Is each item a necessity, or a luxury?

b. For each commodity equation (ii), (iv), and (vi), calculate the expenditure elasticity with respect to total expenditure at the 25th and 75th percentiles of *TOTEXP*.

c. For the budget share equations, (i), (iii), and (v), find the elasticities that are given by $\varepsilon = \dfrac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$ (see Exercise 4.12). Are the changes in elasticities between the two percentiles, noticeable? [A standard log-log expenditure model can be obtained using the data, by creating a dependent variable that is the logarithm of the budget share times total expenditure. That is, for example, $\ln(TELEPHONE) = \ln(PTELEPHONE \times TOTEXP)$.]

4.24 Reconsider the presidential voting data (*fair5*) introduced in Exercises 2.23 and 3.24.

a. Using all the data from 1916 to 2012, estimate the regression model $VOTE = \beta_1 + \beta_2 GROWTH + e$. Based on these estimates, what is the predicted value of *VOTE* in favor of the Democrats in 2012? At the time of the election, a Democrat, Barack Obama, was the incumbent. What is the least squares residual for the 2012 election observation?

b. Estimate the regression in (a) using only data up to 2008. Predict the value of *VOTE* in 2012 using the actual value of *GROWTH* for 2012, which was 1.03%. What is the prediction error in this forecast? Is it larger or smaller than the error computed in part (a).

c. Using the regression results from (b), construct a 95% prediction interval for the 2012 value of *VOTE* using the actual value of $GROWTH = 1.03\%$.

d. Using the estimation results in (b), what value of *GROWTH* would have led to a prediction that the nonincumbent party [Republicans] would have won 50.1% of the vote in 2012?

e. Use the estimates from part (a), and predict the percentage vote in favor of the Democratic candidate in 2016. At the time of the election, a Democrat, Barack Obama, was the incumbent. Choose several values for *GROWTH* that represent both pessimistic and optimistic values for 2016. Cite the source of your chosen values for *GROWTH*.

4.25 The file *collegetown* contains data on 500 houses sold in Baton Rouge, LA during 2009–2013. Variable descriptions are in the file *collegetown.def*.

a. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$. Interpret the estimated model parameters. Calculate the slope and elasticity at the sample means, if necessary.

b. Estimate the log-log model $\ln(PRICE) = \alpha_1 + \alpha_2 \ln(SQFT) + e$. Interpret the estimated parameters. Calculate the slope and elasticity at the sample means, if necessary.

c. Compare the R^2 value from the linear model $PRICE = \delta_1 + \delta_2 SQFT + e$ to the "generalized" R^2 measure for the models in (b) and (c).

d. Construct histograms of the least squares residuals from each of the models in (a)–(c) and obtain the Jarque–Bera statistics. Based on your observations, do you consider the distributions of the residuals to be compatible with an assumption of normality?

e. For each of the models in (a)–(c), plot the least squares residuals against *SQFT*. Do you observe any patterns?

f. For each model in (a)–(c), predict the value of a house with 2700 square feet.

g. For each model in (a)–(c), construct a 95% prediction interval for the value of a house with 2700 square feet.

h. Based on your work in this problem, discuss the choice of functional form. Which functional form would you use? Explain.

4.26 The file *collegetown* contains data on 500 houses sold in Baton Rouge, LA during 2009–2013. Variable descriptions are in the file *collegetown.def.*

a. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$ for houses close to Louisiana State University [$CLOSE = 1$] and again for houses that are not close to Louisiana State University. How similar are the two sets of regression estimates. For each find the "corrected" predictor for a house with 2700 square feet of living area. What do you find?

b. Using the sample of homes that are not close to LSU [$CLOSE = 0$], find any observations on house sales that you would classify as unusual, based on the studentized residuals, *LEVERAGE*, *DFBETAS,* and *DFFITS*. Can you identify any house characteristics that might explain why they are unusual?

c. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$ for houses for which $AGE < 7$ and again for houses with $AGE > 9$. Note that *AGE* is not the actual age of the house, but a category. Examine the file *collegetown.def* for the specifics. How similar are the two sets of regression estimates. For each find the "corrected" predictor of a house with 2700 square feet of living area. What do you find?

d. Using the sample of homes with $AGE > 9$, find any observations on house sales that you would classify as unusual, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*. Can you identify any house characteristics that might explain why they are unusual?

4.27 Does the return to education differ by race and gender? For this exercise use the file *cps5*. [This is a large file with 9799 observations. If your software is a student version, you can use the smaller file *cps5_small* if your instructor permits]. In this exercise, you will extract subsamples of observations consisting of (i) white males, (ii) white females, (iii) black males, and (iv) black females.

a. For each sample partition, obtain the summary statistics of *WAGE*.

b. A variable's **coefficient of variation** (*CV*) is 100 times the ratio of its sample standard deviation to its sample mean. For a variable *y,* it is

$$CV = 100 \times \frac{s_y}{\bar{y}}$$

It is a measure of variation that takes into account the size of the variable. What is the coefficient of variation for *WAGE* within each sample partition?

c. For each sample partition, estimate the log-linear model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$$

What is the approximate percentage return to another year of education for each group?

d. Create 95% interval estimates for the coefficient β_2 in each partition. Identify partitions for which the 95% interval estimates of the rate of return to education *do not* overlap. What does this imply about the population relations between wages and education for these groups? Are they similar or different? For the nonoverlapping pairs, test the null hypothesis that the parameter β_2 in one sample partition (the larger one, for simplicity) equals the estimated value in the other partition, using the 5% level of significance.

e. Create 95% interval estimates for the intercept coefficient in each partition. Identify partitions for which the 95% interval estimates for the intercepts *do not* overlap. What does this imply about the population relations between wages and education for these groups? Are they similar or different? For the nonoverlapping pairs, test the null hypothesis that the parameter β_1 in one sample partition (the larger one, for simplicity) equals the estimated value in the other partition, using the 5% level of significance.

f. Does the model fit the data equally well for each sample partition?

4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

$$
\begin{aligned}
YIELD_t &= \beta_0 + \beta_1 TIME + e_t \\
YIELD_t &= \alpha_0 + \alpha_1 \ln(TIME) + e_t \\
YIELD_t &= \gamma_0 + \gamma_1 TIME^2 + e_t \\
\ln(YIELD_t) &= \phi_0 + \phi_1 TIME + e_t
\end{aligned}
$$

a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

b. Interpret the coefficient of the time-related variable in your chosen specification.

c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.

d. Using your chosen specification, use the observations up to 1996 to estimate the model. Construct a 95% prediction interval for *YIELD* in 1997. Does your interval contain the true value?

4.29 Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three-person households, consisting of a husband, a wife, plus one other. In this exercise, we examine expenditures on a staple item, food. In this extended example, you are asked to compare the linear, log-log, and linear-log specifications.

a. Calculate summary statistics for the variables: *FOOD* and *INCOME*. Report for each the sample mean, median, minimum, maximum, and standard deviation. Construct histograms for both variables. Locate the variable mean and median on each histogram. Are the histograms symmetrical and "bell-shaped" curves? Is the sample mean larger than the median, or vice versa? Carry out the Jarque–Bera test for the normality of each variable.

b. Estimate the linear relationship $FOOD = \beta_1 + \beta_2 INCOME + e$. Create a scatter plot *FOOD* versus *INCOME* and include the fitted least squares line. Construct a 95% interval estimate for β_2. Have we estimated the effect of changing income on average *FOOD* relatively precisely, or not?

c. Obtain the least squares residuals from the regression in (b) and plot them against *INCOME*. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. Is it more important for the variables *FOOD* and *INCOME* to be normally distributed, or that the random error e be normally distributed? Explain your reasoning.

d. Calculate both a point estimate and a 95% interval estimate of the elasticity of food expenditure with respect to income at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Are the estimated elasticities similar or dissimilar? Do the interval estimates overlap or not? As *INCOME* increases should the income elasticity for food increase or decrease, based on Economics principles?

e. For expenditures on food, estimate the log-log relationship $\ln(FOOD) = \gamma_1 + \gamma_2 \ln(INCOME) + e$. Create a scatter plot for $\ln(FOOD)$ versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plot in (b). Is the relationship more or less well-defined for the log-log model

relative to the linear specification? Calculate the generalized R^2 for the log-log model and compare it to the R^2 from the linear model. Which of the models seems to fit the data better?

f. Construct a point and 95% interval estimate of the elasticity for the log-log model. Is the elasticity of food expenditure from the log-log model similar to that in part (d), or dissimilar? Provide statistical evidence for your claim.

g. Obtain the least squares residuals from the log-log model and plot them against ln(*INCOME*). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus ln(*INCOME*) and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to the others? Compare the R^2 values. Which of the models seems to fit the data better?

i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide statistical evidence for your claim.

j. Obtain the least squares residuals from the linear-log model and plot them against ln(*INCOME*). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

4.30 Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three person households, consisting of a husband, a wife, plus one other. In this exercise, we examine expenditures on alcoholic beverages.

a. Obtain summary statistics for *ALCBEV*. How many households spend nothing on alcoholic beverages? Calculate the summary statistics restricting the sample to those households with positive expenditure on alcoholic beverages.

b. Plot *ALCBEV* against *INCOME* and include the fitted least squares regression line. Obtain the least squares estimates of the model $ALCBEV = \beta_1 + \beta_2 INCOME + e$. Obtain the least squares residuals and plot these versus *INCOME*. Does this plot appear random, as in Figure 4.7(a)? If the dependent variable in this regression model is zero ($ALCBEV = 0$), what is the least squares residual? For observations with $ALCBEV = 0$, is the least squares residual related to the explanatory variable *INCOME*? How?

c. Suppose that some households in this sample may never purchase alcohol, regardless of their income. If this is true, do you think that a linear regression including all the observations, even the observations for which $ALCBEV = 0$, gives a reliable estimate of the effect of income on average alcohol expenditure? If there is estimation bias, is the bias positive (the slope overestimated) or negative (slope underestimated)? Explain your reasoning.

d. For households with $ALCBEV > 0$, construct histograms for *ALCBEV* and ln(*ALCBEV*). How do they compare?

e. Create a scatter plot of ln(*ALCBEV*) against ln(*INCOME*) and include a fitted regression line. Interpret the coefficient of ln(*INCOME*) in the estimated log-log regression. How many observations are included in this estimation?

f. Calculate the least squares residuals from the log-log model. Create a histogram of these residuals and also plot them against ln(*INCOME*). Does this plot appear random, as in Figure 4.7(a)?

g. If we consider only the population of individuals who have positive expenditures for alcohol, do you prefer the linear relationship model, or the log-log model?

h. Expenditures on apparel have some similar features to expenditures on alcoholic beverages. You might reconsider the above exercises for *APPAR*. Think about part (c) above. Of those with no apparel expenditure last month, do you think there is a substantial portion who never purchase apparel regardless of income, or is it more likely that they sometimes purchase apparel but simply did not do so last month?

Appendix 4A Development of a Prediction Interval

The forecast error is $f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$. To obtain its variance, let us first obtain the variance of $\hat{y}_0 = b_1 + b_2 x_0$. The variances and covariance of the least squares estimators are given in Section 2.4.4. Using them, and obtaining a common denominator, we obtain

$$\operatorname{var}(\hat{y}_0|\mathbf{x}) = \operatorname{var}\left[(b_1 + b_2 x_0)|\mathbf{x}\right] = \operatorname{var}(b_1|\mathbf{x}) + x_0^2 \operatorname{var}(b_2|\mathbf{x}) + 2x_0 \operatorname{cov}(b_1, b_2|\mathbf{x})$$

$$= \frac{\sigma^2}{N\sum(x_i - \bar{x})^2}\left[\sum x_i^2 + Nx_0^2 - 2N\bar{x}x_0\right]$$

The term in brackets can be simplified. First, factor N from the second and third terms to obtain $\sum x_i^2 + Nx_0^2 - 2N\bar{x}x_0 = \sum x_i^2 + N(x_0^2 - 2\bar{x}x_0)$. Complete the square within the parentheses by adding $\bar{x}^2$, and subtracting $N\bar{x}^2$ to keep the equality. Then the term in brackets is

$$\sum x_i^2 - N\bar{x}^2 + N\left(x_0^2 - 2\bar{x}x_0 + \bar{x}^2\right) = \sum(x_i - \bar{x})^2 + N(x_0 - \bar{x})^2$$

Finally

$$\operatorname{var}(\hat{y}_0|\mathbf{x}) = \sigma^2\left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Taking into account that x_0 and the unknown parameters β_1 and β_2 are not random, you should be able to show that $\operatorname{var}(f|\mathbf{x}) = \operatorname{var}(\hat{y}_0|\mathbf{x}) + \operatorname{var}(e_0) = \operatorname{var}(\hat{y}_0|\mathbf{x}) + \sigma^2$. A little factoring gives the result in (4.4). We can construct a standard normal random variable as

$$\frac{f}{\sqrt{\operatorname{var}(f|\mathbf{x})}} \sim N(0, 1)$$

If the forecast error variance in (4.4) is estimated by replacing σ^2 by its estimator $\hat{\sigma}^2$,

$$\widehat{\operatorname{var}}(f|\mathbf{x}) = \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

then

$$\frac{f}{\sqrt{\widehat{\operatorname{var}}(f|\mathbf{x})}} = \frac{y_0 - \hat{y}_0}{\operatorname{se}(f)} \sim t_{(N-2)} \tag{4A.1}$$

where the square root of the estimated variance is the standard error of the forecast given in (4.5). The t-ratio in (4A.1) is a pivotal statistic. It has a distribution that does not depend on $\mathbf{x}$ or any unknown parameters.

Using these results, we can construct an interval prediction procedure for y_0 just as we constructed confidence intervals for the parameters β_k. If t_c is a critical value from the $t_{(N-2)}$-distribution such that $P(t \geq t_c) = \alpha/2$, then

$$P(-t_c \leq t \leq t_c) = 1 - \alpha \tag{4A.2}$$

Substitute the t-random variable from (4A.1) into (4A.2) to obtain

$$P\left[-t_c \leq \frac{y_0 - \hat{y}_0}{\operatorname{se}(f)} \leq t_c\right] = 1 - \alpha$$

Simplify this expression to obtain

$$P\left[\hat{y}_0 - t_c \operatorname{se}(f) \leq y_0 \leq \hat{y}_0 + t_c \operatorname{se}(f)\right] = 1 - \alpha$$

A $100(1 - \alpha)\%$ confidence interval, or prediction interval, for y_0 is given by (4.6). This prediction interval is valid if $\mathbf{x}$ is fixed or random, as long as assumptions SR1–SR6 hold.

Appendix 4B The Sum of Squares Decomposition

To obtain the sum of squares decomposition in (4.11), we square both sides of (4.10)

$$(y_i - \bar{y})^2 = \left[(\hat{y}_i - \bar{y}) + \hat{e}_i\right]^2 = (\hat{y}_i - \bar{y})^2 + \hat{e}_i^2 + 2(\hat{y}_i - \bar{y})\hat{e}_i$$

Then sum

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2 + 2\sum (\hat{y}_i - \bar{y})\hat{e}_i$$

Expanding the last term, we obtain

$$\sum (\hat{y}_i - \bar{y})\hat{e}_i = \sum \hat{y}_i\hat{e}_i - \bar{y}\sum \hat{e}_i = \sum (b_1 + b_2 x_i)\hat{e}_i - \bar{y}\sum \hat{e}_i$$
$$= b_1 \sum \hat{e}_i + b_2 \sum x_i\hat{e}_i - \bar{y}\sum \hat{e}_i$$

Consider first the term $\sum \hat{e}_i$

$$\sum \hat{e}_i = \sum (y_i - b_1 - b_2 x_i) = \sum y_i - Nb_1 - b_2 \sum x_i = 0$$

This last expression is zero because of the first normal equation (2A.3). The first normal equation is valid *only if the model contains an intercept*. The sum of the least squares residuals is always zero *if* the model contains an intercept. It follows, then, that the *sample mean* of the least squares residuals is also zero (since it is the sum of the residuals divided by the sample size) if the model contains an intercept. That is, $\bar{\hat{e}} = \sum \hat{e}_i / N = 0$.

The next term $\sum x_i\hat{e}_i = 0$, because

$$\sum x_i\hat{e}_i = \sum x_i(y_i - b_1 - b_2 x_i) = \sum x_i y_i - b_1 \sum x_i - b_2 \sum x_i^2 = 0$$

This result follows from the second normal equation (2A.4). This result always holds for the least squares estimator and does not depend on the model having an intercept. See Appendix 2A for discussion of the normal equations. Substituting $\sum \hat{e}_i = 0$ and $\sum x_i\hat{e}_i = 0$ back into the original equation, we obtain $\sum (\hat{y}_i - \bar{y})\hat{e}_i = 0$.

Thus, if the model contains an intercept, it is guaranteed that $SST = SSR + SSE$. If, however, the model does not contain an intercept, then $\sum \hat{e}_i \neq 0$ and $SST \neq SSR + SSE$.

Appendix 4C Mean Squared Error: Estimation and Prediction

In Chapter 2, we discussed the properties of the least squares estimator. Under assumptions SR1–SR5, the least squares estimator is the **B**est **L**inear **U**nbiased **E**stimator (BLUE). There are no estimators that are both linear and unbiased that are better than the least squares estimator. However, this rules out many alternative estimators that statisticians and econometricians have developed over the years, which might be useful in certain contexts. Mean squared error (MSE) is an alternative metric for the quality of an estimator that doesn't depend on linearity or unbiasedness, and hence is more general.

In the linear regression model $y = \beta_1 + \beta_2 x + e$, suppose that we are keenly interested in obtaining an estimate of β_2 that is as close as possible to the true value. The mean squared error of an estimator $\hat{\beta}_2$ is

$$\text{MSE}\left(\hat{\beta}_2\right) = E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right] \tag{4C.1}$$

The term $\left(\hat{\beta}_2 - \beta_2\right)^2$ is the squared estimation error, that is, the squared difference or distance between the estimator $\hat{\beta}_2$ and the parameter β_2 of interest. Because the estimator $\hat{\beta}_2$ exhibits sampling variation, it is a random variable, and the squared term $\left(\hat{\beta}_2 - \beta_2\right)^2$ is also random. If we

think of "expected value" as "the average in all possible samples," then the mean squared error $E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right]$ is the average, or mean, squared error using $\hat{\beta}_2$ as an estimator of β_2. It measures how close the estimator $\hat{\beta}_2$ is on average to the true parameter β_2. We would like an estimator that is as close as possible to the true parameter and one that has a small mean squared error.

An interesting feature of an estimator's mean squared error is that it takes into account both the estimator's bias and its sampling variance. To see this we play a simple trick on equation (4C.1); we will add and subtract $E\left(\hat{\beta}_2\right)$ inside the parentheses and then square the result. That is,

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\beta}_2\right) &= E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right] = E\left\{\left(\hat{\beta}_2 \underbrace{- E\left(\hat{\beta}_2\right) + E\left(\hat{\beta}_2\right)}_{=0} - \beta_2\right)^2\right\} \\
&= E\left\{\left(\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right] + \left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right)^2\right\} \\
&= E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]^2\right\} + E\left\{\left[E\left(\hat{\beta}_2\right) - \beta_2\right]^2\right\} \\
&\quad + 2E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right\} \\
&= \mathrm{var}\left(\hat{\beta}_2\right) + \left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2
\end{aligned} \tag{4C.2}$$

To go from the third to the fourth lines, we first recognize that $E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]^2\right\} = \mathrm{var}\left(\hat{\beta}_2\right)$. Secondly, in the term $E\left\{\left[E\left(\hat{\beta}_2\right) - \beta_2\right]^2\right\}$, the outside expectation is not needed because $E\left(\hat{\beta}_2\right)$ is not random and β_2 is not random. The difference between an estimator's expected value and the true parameter is called the **estimator bias**, so $E\left(\hat{\beta}_2\right) - \beta_2 = \mathrm{bias}\left(\hat{\beta}_2\right)$. The term $\left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2$ is the squared estimator bias. The final term in the third line of (4C.2) is zero. To see this note again that $\left[E\left(\hat{\beta}_2\right) - \beta_2\right]$ is not random, so that it can be factored out of the expectation

$$\begin{aligned}
2E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right\} &= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\left\{E\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\right\} \\
&= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\left[E\left(\hat{\beta}_2\right) - E\left(\hat{\beta}_2\right)\right] \\
&= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]0 = 0
\end{aligned}$$

We have shown that an estimator's mean squared error is the sum of its variance and squared bias,

$$\mathrm{MSE}\left(\hat{\beta}_2\right) = \mathrm{var}\left(\hat{\beta}_2\right) + \left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2 \tag{4C.3}$$

This relationship is also true if we use conditional expectations. The conditional MSE is

$$\mathrm{MSE}\left(\hat{\beta}_2|\mathbf{x}\right) = \mathrm{var}\left(\hat{\beta}_2|\mathbf{x}\right) + \left[\mathrm{bias}\left(\hat{\beta}_2|\mathbf{x}\right)\right]^2 \tag{4C.4}$$

with $\mathrm{bias}\left(\hat{\beta}_2|\mathbf{x}\right) = E\left(\hat{\beta}_2|\mathbf{x}\right) - \beta_2$. Because the least squares estimator is unbiased under SR1–SR5, its mean squared error is

$$\mathrm{MSE}\left(b_2|\mathbf{x}\right) = \mathrm{var}\left(b_2|\mathbf{x}\right) + \left[\mathrm{bias}\left(b_2|\mathbf{x}\right)\right]^2 = \mathrm{var}\left(b_2|\mathbf{x}\right) + [0]^2 = \mathrm{var}\left(b_2|\mathbf{x}\right) \tag{4C.5}$$

The mean squared error concept can also be applied to more than one parameter at once. For example, the mean squared error of $\hat{\beta}_1$ and $\hat{\beta}_2$ as estimators of β_1 and β_2 is

$$\begin{aligned}\text{MSE}\left(\hat{\beta}_1, \hat{\beta}_2|\mathbf{x}\right) &= E\left\{\left[\left(\hat{\beta}_1-\beta_1\right)^2+\left(\hat{\beta}_2-\beta_2\right)^2\right]\middle|\mathbf{x}\right\}\\ &= \text{var}\left(\hat{\beta}_1|\mathbf{x}\right)+\left[\text{bias}\left(\hat{\beta}_1|\mathbf{x}\right)\right]^2+\text{var}\left(\hat{\beta}_2|\mathbf{x}\right)+\left[\text{bias}\left(\hat{\beta}_2|\mathbf{x}\right)\right]^2\end{aligned}$$

In the simple linear regression model, there are no estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β_1 and β_2 that have mean squared error $\text{MSE}\left(\hat{\beta}_1, \hat{\beta}_2|\mathbf{x}\right)$ smaller than the mean squared error for the least squares estimator, $\text{MSE}\left(b_1, b_2|\mathbf{x}\right)$, for any and all parameter values. This statement turns out not to be true in the multiple regression model.

We can apply the mean squared error concept to prediction situations too. Suppose that we are predicting an outcome y_0 using a predictor $\hat{y}_0(\mathbf{x})$, which is a function of the sample $\mathbf{x}$. The conditional mean squared error of the predictor is $E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]$. We employ the same trick as in (4C.2), adding and subtracting $E\left(y_0|\mathbf{x}\right)$, the conditional expected value of y_0,

$$\begin{aligned}E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right] &= E\left[\left(y_0-E\left(y_0|\mathbf{x}\right)+E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]\\ &= E\left[\left(y_0-E\left(y_0|\mathbf{x}\right)\right)^2\middle|\mathbf{x}\right]+E\left[\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]\\ &\quad+2E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]\right)\middle|\mathbf{x}\right\}\\ &= \text{var}\left(y_0|\mathbf{x}\right)+\left\{\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]^2\middle|\mathbf{x}\right\}\end{aligned} \tag{4C.6}$$

The third line in (4C.6) is zero because conditional on $\mathbf{x}$ the term $E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})$ is not random, and it can be factored out of the expectation

$$\begin{aligned}2E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]\right)\middle|\mathbf{x}\right\} &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\right)\middle|\mathbf{x}\right\}\\ &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)\left[E\left(y_0|\mathbf{x}\right)-E\left(y_0|\mathbf{x}\right)\right]\\ &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)\times 0=0\end{aligned}$$

The conditional mean squared error of our predictor is then

$$E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]=\text{var}\left(y_0|\mathbf{x}\right)+\left[\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right] \tag{4C.7}$$

Using the law of iterated expectations

$$E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)\right]^2=E_{\mathbf{x}}\left[\text{var}\left(y_0|\mathbf{x}\right)\right]+E_{\mathbf{x}}\left\{\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]^2\right\} \tag{4C.8}$$

If we are choosing a predictor, then the one that minimizes the mean squared error is $\hat{y}_0(\mathbf{x})=E\left(y_0|\mathbf{x}\right)$. This makes the final term in (4C.8) zero. The conditional mean of y_0 is the minimum mean squared error predictor of y_0.

CHAPTER 5

The Multiple Regression Model

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Recognize a multiple regression model and be able to interpret the coefficients in that model.
2. Understand and explain the meanings of the assumptions for the multiple regression model.
3. Use your computer to find least squares estimates of the coefficients in a multiple regression model, and interpret those estimates.
4. Explain the meaning of the Gauss–Markov theorem.
5. Compute and explain the meaning of R^2 in a multiple regression model.
6. Explain the Frisch–Waugh–Lovell Theorem and estimate examples to show how it works.
7. Use your computer to obtain variance and covariance estimates, and standard errors, for the estimated coefficients in a multiple regression model.
8. Explain the circumstances under which coefficient variances (and standard errors) are likely to be relatively high, and those under which they are likely to be relatively low.
9. Find interval estimates for single coefficients and linear combinations of coefficients, and interpret the interval estimates.
10. Test hypotheses about single coefficients and about linear combinations of coefficients in a multiple regression model. In particular,
    a. What is the difference between a one-tail and a two-tail test?
    b. How do you compute the *p*-value for a one-tail test, and for a two-tail test?
    c. What is meant by "testing the significance of a coefficient"?
    d. What is the meaning of the *t*-values and *p*-values that appear in your computer output?
    e. How do you compute the standard error of a linear combination of coefficient estimates?
11. Estimate and interpret multiple regression models with polynomial and interaction variables.
12. Find point and interval estimates and test hypotheses for marginal effects in polynomial regressions and models with interaction variables.
13. Explain the difference between finite and large sample properties of an estimator.
14. Explain what is meant by consistency and asymptotic normality.

15. Describe the circumstances under which we can use the finite sample properties of the least squares estimator, and the circumstances under which asymptotic properties are required.

16. Use your computer to compute the standard error of a nonlinear function of estimators. Use that standard error to find interval estimates and to test hypotheses about nonlinear functions of coefficients.

KEYWORDS

asymptotic normality
BLU estimator
consistency
covariance matrix of least squares estimators
critical value
delta method
error variance estimate
error variance estimator
explained sum of squares
FWL theorem
goodness-of-fit
interaction variable
interval estimate
least squares estimates
least squares estimation
least squares estimators
linear combinations
marginal effect
multiple regression model
nonlinear functions
one-tail test
p-value
polynomial
regression coefficients
standard errors
sum of squared errors
sum of squares due to regression
testing significance
total sum of squares
two-tail test

The model in Chapters 2–4 is called a simple regression model because the dependent variable y is related to only *one* explanatory variable x. Although this model is useful for a range of situations, in most economic models there are two or more explanatory variables that influence the dependent variable y. For example, in a demand equation the quantity demanded of a commodity depends on the price of that commodity, the prices of substitute and complementary goods, and income. Output in a production function will be a function of more than one input. Aggregate money demand will be a function of aggregate income and the interest rate. Investment will depend on the interest rate and on changes in income.

When we turn an economic model with more than one explanatory variable into its corresponding econometric model, we refer to it as a **multiple regression model**. Most of the results we developed for the simple regression model in Chapters 2–4 can be extended naturally to this general case. There are slight changes in the interpretation of the β parameters, the degrees of freedom for the t-distribution will change, and we will need to modify the assumption concerning the characteristics of the explanatory (x) variables. These and other consequences of extending the simple regression model to a multiple regression model are described in this chapter.

As an example for introducing and analyzing the multiple regression model, we begin with a model used to explain sales revenue for a fast-food hamburger chain with outlets in small U.S. cities.

5.1 Introduction

5.1.1 The Economic Model

We will set up an economic model for a hamburger chain that we call Big Andy's Burger Barn.[1] Important decisions made by the management of Big Andy's include its pricing policy for different products and how much to spend on advertising. To assess the effect of different price

[1]The data we use reflect a real fast-food franchise whose identity we disguise under the name Big Andy's.

structures and different levels of advertising expenditure, Big Andy's Burger Barn sets different prices, and spends varying amounts on advertising, in different cities. Of particular interest to management is how sales revenue changes as the level of advertising expenditure changes. Does an increase in advertising expenditure lead to an increase in sales? If so, is the increase in sales sufficient to justify the increased advertising expenditure? Management is also interested in pricing strategy. Will reducing prices lead to an increase or decrease in sales revenue? If a reduction in price leads only to a small increase in the quantity sold, sales revenue will fall (demand is price-inelastic); a price reduction that leads to a large increase in quantity sold will produce an increase in revenue (demand is price-elastic). This economic information is essential for effective management.

The first step is to set up an economic model in which sales revenue depends on one or more explanatory variables. We initially hypothesize that sales revenue is linearly related to price and advertising expenditure. The economic model is

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT \tag{5.1}$$

where *SALES* represents monthly sales (total) revenue in a given city, *PRICE* represents price in that city, and *ADVERT* is monthly advertising expenditure in that city. Both *SALES* and *ADVERT* are measured in terms of thousands of dollars. Because sales in bigger cities will tend to be greater than sales in smaller cities, we focus on smaller cities with comparable populations.

Since a hamburger outlet sells a number of products—burgers, fries, and shakes—and each product has its own price, it is not immediately clear what price should be used in (5.1). What we need is some kind of average price for all products and information on how this average price changes from city to city. For this purpose, management has constructed a single price index *PRICE*, measured in dollars and cents, that describes overall prices in each city.

The remaining symbols in (5.1) are the unknown parameters β_1, β_2, and β_3 that describe the dependence of sales (*SALES*) on price (*PRICE*) and advertising (*ADVERT*). To be more precise about the interpretation of these parameters, we move from the economic model in (5.1) to an econometric model that makes explicit assumptions about the way the data are generated.

5.1.2 The Econometric Model

When we collect data on *SALES*, *PRICE*, and *ADVERT* from the franchises in different cities, the observations will not exactly satisfy the linear relationship described in equation (5.1). The behavior of Andy's customers in different cities will not be such that the same prices and the same level of advertising expenditure will always lead to the same sales revenue. Other factors not in the equation likely to affect sales include the number and behavior of competing fast-food outlets, the nature of the population in each city—their age profile, income, and food preferences—and the location of Andy's burger barns—near a busy highway, downtown, and so on. To accommodate these factors, we include an error term e in the equation so that the model becomes

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e \tag{5.2}$$

As discussed in Chapter 2, the way in which data are collected has a bearing on what assumptions are relevant and realistic for the error term e, the explanatory variables *PRICE* and *ADVERT*, and the dependent variable *SALES*. These assumptions in turn affect how we make inferences about the parameters β_1, β_2, and β_3.

Assume we take a random sample of 75 franchises in similar-sized cities in which Big Andy operates, and we observe their monthly sales, prices, and advertising expenditure. Thus, we have observations $(SALES_i, PRICE_i, ADVERT_i)$ for $i = 1, 2, \ldots, 75$. Because we do not know which cities will be chosen before we randomly sample, the triplet $(SALES_i, PRICE_i, ADVERT_i)$ is a three-dimensional random variable with a joint probability distribution. Also, the fact that we have a **random** sample implies that the observations from different cities are independent. That is,

$(SALES_i, PRICE_i, ADVERT_i)$ is independent of $(SALES_j, PRICE_j, ADVERT_j)$ for $i \neq j$. Associated with each observation is another random variable, the unobservable error term e_i that reflects the effect of factors other than *PRICE* and *ADVERT* on *SALES*. The model for the *i*th observation is written as

$$SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i \tag{5.3}$$

We assume that the effect of e_i on sales, averaged over all cities in the population, is zero, and that knowing *PRICE* and *ADVERT* for a given city does not help us predict the value of *e* for that city. At each $(PRICE_i, ADVERT_i)$ pair of values the average of the random errors is zero, that is,

$$E(e_i | PRICE_i, ADVERT_i) = 0 \tag{5.4}$$

This assumption, when combined with the assumption of independent observations generated from a random sample, implies that e_i is **strictly exogenous**. How do we check whether this is a reasonable assumption? We need to ask whether e_i includes any variables that have an effect on *SALES* (are correlated with *SALES*), *and* are also correlated with *PRICE* or *ADVERT*. If the answer is yes, strict exogeneity is violated. This might happen, for example, if the pricing and advertising behavior of Andy's competitors affects his sales, and is correlated with his own pricing and advertising policies. At the moment, it is convenient if we abstract from such a situation and continue with the strict exogeneity assumption.[2]

Using equations (5.3) and (5.4), we can write

$$E(SALES|PRICE, ADVERT) = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT \tag{5.5}$$

Equation (5.5) is the *conditional mean* or *conditional expectation* of *SALES* given *PRICE* and *ADVERT* and is known as the **multiple regression function** or simply the **regression function**. It shows how the population average or population mean value for *SALES* changes depending on the settings for price and advertising expenditure. For given values of *PRICE* and *ADVERT*, some *SALES* values will fall above the mean and some below. We have dropped the subscript *i* for convenience and to emphasize that we assume this relationship holds for all cities in the population.

With this background, how do we interpret each of the parameters β_1, β_2, and β_3? Mathematically, the intercept parameter β_1 is the expected value of the dependent variable when each of the independent, explanatory variables takes the value zero. However, in many cases this parameter has no clear economic interpretation. In this particular case, it is not realistic to have a situation in which $PRICE = ADVERT = 0$. Except in very special circumstances, we always include an intercept in the model, even if it has no direct economic interpretation. Omitting it can lead to a model that fits the data poorly and that does not predict well.

The other parameters in the model measure the change in the expected value of the dependent variable given a unit change in an explanatory variable, *all other variables held constant*.

β_2 = the change in expected monthly *SALES* ($1000) when the price index *PRICE* is increased by one unit ($1), and advertising expenditure *ADVERT* is held constant

$$= \left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta PRICE}\right|_{(ADVERT \text{ held constant})} = \frac{\partial E(SALES|PRICE, ADVERT)}{\partial PRICE}$$

The symbol "∂" stands for "partial differentiation." Those of you familiar with calculus may have seen this operation. In the context above, the partial derivative of average *SALES* with respect to

[2]How to cope with violations of this assumption is considered in Chapter 10.

PRICE is the rate of change of average *SALES* as *PRICE* changes, with other factors, in this case *ADVERT*, held constant. Further details can be found in Appendix A.3.5. We will occasionally use partial derivatives, but not to an extent that will disadvantage you if you have not had a course in calculus. Rules for differentiation are provided in Appendix A.3.1.

The sign of β_2 could be positive or negative. If an increase in price leads to an increase in sales revenue, then $\beta_2 > 0$, and the demand for the chain's products is price-inelastic. Conversely, a price-elastic demand exists if an increase in price leads to a decline in revenue, in which case $\beta_2 < 0$. Thus, knowledge of the *sign* of β_2 provides information on the price-elasticity of demand. The *magnitude* of β_2 measures the amount of change in revenue for a given price change.

The parameter β_3 describes the response of expected sales revenue to a change in the level of advertising expenditure. That is,

$$\beta_3 = \text{the change in expected monthly } SALES(\$1000) \text{ when advertising expenditure } ADVERT \text{ is increased by one unit (\$1000), and the price index } PRICE \text{ is held constant}$$

$$= \left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta ADVERT}\right|_{(PRICE \text{ held constant})} = \frac{\partial E(SALES|PRICE, ADVERT)}{\partial ADVERT}$$

We expect the sign of β_3 to be positive. That is, we expect that an increase in advertising expenditure, unless the advertising is offensive, will lead to an increase in sales revenue. Whether or not the increase in revenue is sufficient to justify the added advertising expenditure, as well as the added cost of producing more hamburgers, is another question. With $\beta_3 < 1$, an increase of \$1000 in advertising expenditure will yield an increase in revenue that is less than \$1000. For $\beta_3 > 1$, it will be greater. Thus, in terms of the chain's advertising policy, knowledge of β_3 is very important.

Critical to the above interpretations for β_2 and β_3 is the strict exogeneity assumption $E(e_i|PRICE_i, ADVERT_i) = 0$. It implies that β_2, for example, can be interpreted as the effect of *PRICE* on *SALES*, holding all other factors constant, including the unobservable factors that form part of the error term *e*. We can say that a one-unit change in *PRICE causes* mean *SALES* to change by β_2 units. If the exogeneity assumption does not hold, the parameters cannot be given this causal interpretation. When $E(e_i|PRICE_i) \neq 0$, a change in price is correlated with the error term and hence the effect of a change in price cannot be captured by β_2 alone. For example, suppose that Big Andy's main competitor is Little Jim's Chicken House. And suppose that every time Andy changes his burger price, Jim responds by changing his chicken price. Because Jim's chicken price is not explicitly included in the equation, but is likely to impact on Andy's sales, its effect will be included in the error term. Also, because Jim's price is correlated with Andy's price, $E(e_i|PRICE_i) \neq 0$. Thus, a change in Andy's price (*PRICE*) will impact on *SALES* through both β_2 and the error term. Note, however, if Jim's price is added to the equation as another variable, instead of forming part of the error term, and the new error term satisfies the exogeneity assumption, then the causal interpretation of the parameter is retained.

Similar remarks can be made about the parameter for *ADVERT*, β_3.

EXAMPLE 5.1 | Data for Hamburger Chain

In the simple regression model in Chapters 2–4, the regression function was represented graphically by a line describing the relationship between $E(y|x)$ and x. With the multiple regression model with two explanatory variables, equation (5.5) describes not a line but a *plane*. As illustrated in Figure 5.1, the plane intersects the vertical axis at β_1. The parameters β_2 and β_3 measure the slope of the plane in the directions of the "price axis" and the "advertising

axis," respectively. Representative observations for sales revenue, price, and advertising for some cities are displayed in Table 5.1. The complete set of observations can be found in the data file *andy* and is represented by the dots in Figure 5.1. These data do not fall exactly on a plane but instead resemble a "cloud."

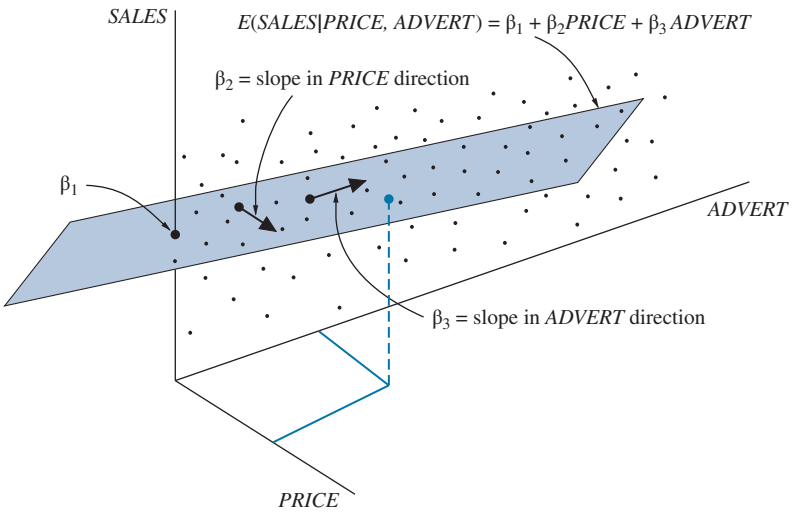


FIGURE 5.1 **The multiple regression plane.**

TABLE 5.1 **Observations on Monthly Sales, Price, and Advertising**

City	*SALES* $1000 units	*PRICE* $1 units	*ADVERT* $1000 units
1	73.2	5.69	1.3
2	71.8	6.49	2.9
3	62.4	5.63	0.8
4	67.4	6.22	0.7
5	89.3	5.02	1.5
.	.	.	.
.	.	.	.
.	.	.	.
73	75.4	5.71	0.7
74	81.3	5.45	2.0
75	75.0	6.05	2.2
	Summary statistics		
Sample mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Maximum	91.20	6.49	3.10
Minimum	62.40	4.83	0.50
Std. Dev.	6.49	0.52	0.83

5.1.3 The General Model

It is useful to digress for a moment and summarize how the concepts developed so far relate to the general case. Working in this direction, let

$$y_i = SALES_i \quad x_{i2} = PRICE_i \quad x_{i3} = ADVERT_i$$

Then, equation (5.3) can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \tag{5.6}$$

You might wonder why we have defined x_{i2} and x_{i3}, and not x_{i1}. We can think of the first term on the right-hand side of the equation as $\beta_1 x_{i1}$ where $x_{i1} = 1$, that is, x_{i1} is equal to 1 for all observations; it is called the **constant term**.

In Chapter 2, we used the notation $\mathbf{x}$ to denote all sample observations on a single variable x. Now that we have observations on two explanatory variables, we use the notation $\mathbf{X}$ to denote all observations on both variables as well as the constant term x_{i1}. That is, $\mathbf{X} = \{(1, x_{i2}, x_{i3}), i = 1, 2, \ldots, N\}$. In the Burger Barn example, $N = 75$. Also, it will sometimes be convenient to denote the ith observation as $\mathbf{x}_i = (1, x_{i2}, x_{i3})$. Given this setup, the strict exogeneity assumption for the Burger Barn example, where we have a random sample with independent $\mathbf{x}_i$, is $E(e_i|\mathbf{x}_i) = 0$. For more general data generating processes where the different sample observations on $\mathbf{x}_i$ are correlated with each other, the strict exogeneity assumption is written as $E(e_i|\mathbf{X}) = 0$. If you need a refresher on the difference between $E(e_i|\mathbf{x}_i) = 0$ and $E(e_i|\mathbf{X}) = 0$, please go back and reread Section 2.2. Correlation between different observations (different $\mathbf{x}_i$) typically exists when using time-series data. In the Burger Barn example, it could occur if our sample was not random, but taken as a collection of Barns from each of a number of states, and the pricing-advertising policies were similar for all Barns within a particular state.

We have noted the implications of the strict exogeneity assumption for the interpretation of the parameters β_2 and β_3. Later, we discuss the implications for estimator properties and inference.

There are many multiple regression models where we have more than two explanatory variables. For example, the Burger Barn model could include the price of Little Jim's Chicken, and an indicator variable equal to 1 if a Barn is near a major highway interchange, and zero otherwise. The ith observation for the general model with $K - 1$ explanatory variables and a constant term can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$$

The definitions of $\mathbf{X}$ and $\mathbf{x}_i$ extend readily to this general case with $\mathbf{X} = \{(1, x_{i2}, \ldots, x_{iK}), i = 1, 2, \ldots, N\}$ and $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iK})$. If strict exogeneity $E(e_i|\mathbf{X}) = 0$ holds, the multiple regression function is

$$E(y_i|\mathbf{X}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} \tag{5.7}$$

The unknown parameters $\beta_2, \beta_3, \ldots, \beta_K$ correspond to the explanatory variables $x_2, x_3, \ldots, x_K$. Because of this correspondence, we will also refer to $\beta_2, \beta_3, \ldots, \beta_K$ as the **coefficients** of $x_2, x_3, \ldots, x_K$. A single coefficient, call it β_k, measures the effect of a change in the variable x_k upon the expected value of y, all other variables held constant. In terms of partial derivatives,

$$\beta_k = \left.\frac{\Delta E(y|x_2, x_3, \ldots, x_K)}{\Delta x_k}\right|_{\text{other } x\text{'s held constant}} = \frac{\partial E(y|x_2, x_3, \ldots, x_K)}{\partial x_k}$$

The parameter β_1 is the intercept term. We use K to denote the total number of unknown parameters in (5.7). For a large part of this chapter, we will introduce point and interval estimation in terms of the model with $K = 3$. The results generally hold for models with more explanatory variables ($K > 3$).

5.1.4 Assumptions of the Multiple Regression Model

To complete our specification of the multiple regression model, we make further assumptions about the error term and the explanatory variables. These assumptions align with those made for the simple regression model in Section 2.2. Their purpose is to establish a framework for estimating the unknown parameters β_k, deriving the properties of the estimator for the β_k, and testing hypotheses of interest about those unknown coefficients. As we travel through the book, we discover that some of the assumptions are too restrictive for some samples of data, requiring us to weaken many of the assumptions. We will examine the implications of changes to the assumptions for estimation and hypothesis testing.

MR1: Econometric Model Observations on $(y_i, \mathbf{x}_i) = (y_i, x_{i2}, x_{i3}, \ldots x_{iK})$ satisfy the population relationship

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$$

MR2: Strict Exogeneity The conditional expectation of the random error e_i, given all explanatory variable observations $\mathbf{X} = \{\mathbf{x}_i, i = 1, 2, \ldots, N\}$, is zero.

$$E(e_i|\mathbf{X}) = 0$$

This assumption implies $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0$ for $k = 1, 2, \ldots, K$ and $(i, j) = 1, 2, \ldots, N$. Each random error has a probability distribution with zero mean. Some errors will be positive, some will be negative; over a large number of observations they will average out to zero. Also, all the explanatory variables are uncorrelated with the error; knowing values of the explanatory variables does not help predict the value of e_i. Thus, the observations will be scattered evenly above and below a plane like the one depicted in Figure 5.1. Fitting a plane through the data will make sense. Another implication of the strict exogeneity assumption is that the multiple regression function is given by

$$E(y_i|\mathbf{X}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK}$$

The mean of the conditional distribution of the dependent variable y_i is a linear function of the explanatory variables $\mathbf{x}_i = (x_{i2}, x_{i3}, \ldots, x_{iK})$.

MR3: Conditional Homoskedasticity The variance of the error term, conditional on $\mathbf{X}$, is a constant.

$$\text{var}(e_i|\mathbf{X}) = \sigma^2$$

This assumption implies $\text{var}(y_i|\mathbf{X}) = \sigma^2$ is a constant. The variability of y_i around its conditional mean function $E(y_i|\mathbf{X}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK}$ does not depend on $\mathbf{X}$. The errors are not more or less likely to be larger for some values of the explanatory variables than for others. Errors with this property are said to be *homoskedastic*.[3]

MR4: Conditionally Uncorrelated Errors The covariance between different error terms e_i and e_j, conditional on $\mathbf{X}$, is zero.

$$\text{cov}(e_i, e_j|\mathbf{X}) = 0 \quad \text{for} \quad i \neq j$$

[3] Because $E(e_i|\mathbf{X}) = 0$, the unconditional variance of e_i is also constant. That is, $\text{var}(e_i) = \sigma^2$. We cannot make the same statement about the unconditional variance for y_i, however. See Appendix B, equation (B.27) for the relationship between conditional and unconditional variances.

All pairs of errors are uncorrelated. The covariance between two random errors corresponding to any two different observations is zero for all values of $\mathbf{X}$. There is no covariation or co-movement in the errors in the sense that the size of an error for one observation has no bearing on the likely size of an error for another observation. With cross-sectional data, this assumption implies that there is no spatial correlation between the errors. With time-series data, it implies there is no correlation in the errors over time. When it exists, correlation over time is referred to as serial or autocorrelation. We typically use subscripts t and s with time-series data and hence the assumption of no serial correlation can be written alternatively as $\text{cov}(e_t, e_s|\mathbf{X}) = 0$ for $t \neq s$.[4]

MR5: No Exact Linear Relationship Exists Between the Explanatory Variables It is not possible to express one of the explanatory variables as an exact linear function of the others. Mathematically, we write this assumption as saying: The only values of $c_1, c_2, \ldots, c_K$ for which

$$c_1 x_{i1} + c_2 x_{i2} + \cdots + c_K x_{iK} = 0 \quad \text{for all observations} \quad i = 1, 2, \ldots, N \tag{5.8}$$

are the values $c_1 = c_2 = \cdots = c_K = 0$. If (5.8) holds and one or more of the c_k's can be nonzero, the assumption is violated. To appreciate why this assumption is necessary, it is useful to consider some special case violations. First, suppose $c_2 \neq 0$ and the other c_k are zero. Then, (5.8) implies $x_{i2} = 0$ for all observations. If $x_{i2} = 0$, then we cannot hope to estimate β_2, which measures the effect of a change in x_{i2} on y_i, with all other factors held constant. As a second special case, suppose c_2, c_3, and c_4 are nonzero and the other c_k are zero. Then, from (5.8) we can write $x_{i2} = -(c_3/c_2)\, x_{i3} - (c_4/c_2)\, x_{i4}$. In this case, x_{i2} is an exact linear function of x_{i3} and x_{i4}. This relationship presents problems because changes in x_{i2} are completely determined by changes in x_{i3} and x_{i4}. It is not possible to separately estimate the effects of changes in each of these three variables. Put another way, there is no independent variation in x_{i2} that will enable us to estimate β_2. Our third special case relates to assumption SR5 of the simple regression model, which stated that the explanatory variable must vary. Condition (5.8) includes this case. Suppose that there is no variation in x_{i3} such that we can write $x_{i3} = 6$ for all i. Then, recalling that $x_{i1} = 1$, we can write $6x_{i1} = x_{i3}$. This outcome violates (5.8), with $c_1 = 6, c_3 = -1$ and the other c_k equal to zero.

MR6: Error Normality (*optional*) Conditional on $\mathbf{X}$, the errors are normally distributed

$$e_i|\mathbf{X} \sim N(0, \sigma^2)$$

This assumption implies that the conditional distribution of y is also normally distributed, $y_i|\mathbf{X} \sim N(E(y_i|\mathbf{X}), \sigma^2)$. It is useful for hypothesis testing and interval estimation when samples are relatively small. However, we call it optional for two reasons. First, it is not necessary for many of the good properties of the least squares estimator to hold. Second, as we will see, if samples are relatively large, it is no longer a necessary assumption for hypothesis testing and interval estimation.

Other Assumptions In the more advanced material in Section 2.10, we considered stronger sets of assumptions for the simple regression model that are relevant for some data generating processes—nonrandom x's, random and independent x, and random sampling, as well as the random and strictly exogenous x case considered here. The properties and characteristics of our inference procedures—estimation and hypothesis testing—established for the random and strictly exogenous x case carry over to cases where stronger assumptions are applicable.

[4] In a similar way to the assumption about conditional homoskedasticity, we can show that $\text{cov}(e_i, e_j|\mathbf{X}) = 0$ implies $\text{cov}(y_i, y_j|\mathbf{X}) = 0$ and $\text{cov}(e_i, e_j) = 0$, but the unconditional covariance $\text{cov}(y_i, y_j)$ may not be zero.

5.2 Estimating the Parameters of the Multiple Regression Model

In this section, we consider the problem of using the least squares principle to estimate the unknown parameters of the multiple regression model. We will discuss estimation in the context of the model in (5.6), which we repeat here for convenience, with i denoting the ith observation.

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

This model is simpler than the full model, yet all the results we present carry over to the general case with only minor modifications.

5.2.1 Least Squares Estimation Procedure

To find an estimator for estimating the unknown parameters we follow the least squares procedure that was first introduced in Chapter 2 for the simple regression model. With the least squares principle, we find those values of $(\beta_1, \beta_2, \beta_3)$ that minimize the sum of squared differences between the observed values of y_i and their expected values $E(y_i|\mathbf{X}) = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3$. Mathematically, we minimize the sum of squares function $S(\beta_1, \beta_2, \beta_3)$, which is a function of the unknown parameters, given the data

$$\begin{aligned} S(\beta_1, \beta_2, \beta_3) &= \sum_{i=1}^{N} (y_i - E(y_i|\mathbf{X}))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3})^2 \end{aligned} \tag{5.9}$$

Given the sample observations y_i, and $\mathbf{x}_i$, minimizing the sum of squares function is a straightforward exercise in calculus. Details of this exercise are given in Appendix 5A. The solutions give us formulas for the least squares estimators for the β coefficients in a multiple regression model with two explanatory variables. They are extensions of those given in (2.7) and (2.8) for the simple regression model with one explanatory variable. There are three reasons for relegating these formulas to Appendix 5A instead of inflicting them on you here. First, they are complicated formulas that we do not expect you to memorize. Second, we never use these formulas explicitly; computer software uses the formulas to calculate least squares estimates. Third, we frequently have models with more than two explanatory variables, in which case the formulas become even more complicated. If you proceed with more advanced study in econometrics, you will discover that there is one relatively simple matrix algebra expression for the least squares estimator that can be used for all models, irrespective of the number of explanatory variables.

Although we always get the computer to do the work for us, it is important to understand the least squares principle and the difference between least squares estimators and least squares estimates. Looked at as a general way to use sample data, formulas for b_1, b_2, and b_3, obtained by minimizing (5.9), are estimation procedures, which are called the **least squares estimators** of the unknown parameters. In general, since their values are not known until the data are observed and the estimates calculated, the least squares estimators are random variables. Computer software applies the formulas to a specific sample of data producing **least squares estimates**, which are numeric values. These least squares estimators and estimates are also referred to as **ordinary** least squares estimators and estimates, abbreviated OLS, to distinguish them from other estimators and estimates such as weighted least squares and two-stage least squares that we encounter later in the book. To avoid too much notation, we use b_1, b_2, and b_3 to denote both the estimators and the estimates.

EXAMPLE 5.2 | OLS Estimates for Hamburger Chain Data

Table 5.2 contains the least squares results for the sales equation for Big Andy's Burger Barn. The least squares estimates are

$$b_1 = 118.91 \quad b_2 = -7.908 \quad b_3 = 1.863$$

Following Example 4.3, these estimates along with their standard errors and the equation's R^2 are typically reported in equation format as

$$\begin{aligned} \widehat{SALES} &= 118.91 - 7.908\,PRICE + 1.863\,ADVERT \\ (\text{se}) &\quad (6.35) \quad (1.096) \quad (0.683) \\ R^2 &= 0.448 \end{aligned} \tag{5.10}$$

From the information in this equation, one can readily construct interval estimates or test hypotheses for each of the β_k in a manner similar to that described in Chapter 3, but with a change in the number of degrees of freedom for the t-distribution. Like before, the t-values and p-values in Table 5.2 relate to testing $H_0: \beta_k = 0$ against the alternative $H_1: \beta_k \neq 0$ for $k = 1, 2, 3$.

We proceed by first interpreting the estimates in (5.10). Then, to explain the degrees of freedom change that arises from having more than one explanatory variable, and to reinforce earlier material, we go over the sampling properties of the least squares estimator, followed by interval estimation and hypothesis testing.

What can we say about the coefficient estimates in (5.10)?

1. The negative coefficient on *PRICE* suggests that demand is price elastic; we estimate that, with advertising held constant, an increase in price of \$1 will lead to a fall in mean monthly revenue of \$7908. Or, expressed differently, a reduction in price of \$1 will lead to an increase in mean revenue of \$7908. If such is the case, a strategy of price reduction through the offering of specials would be successful in increasing sales revenue. We do need to consider carefully the magnitude of the price change, however. A \$1 change in price is a relatively large change. The sample mean of price is 5.69 and its standard deviation is 0.52. A 10-cent change is more realistic, in which case we estimate the mean revenue change to be \$791.
2. The coefficient on advertising is positive; we estimate that with price held constant, an increase in advertising expenditure of \$1000 will lead to an increase in mean sales revenue of \$1863. We can use this information, along with the costs of producing the additional hamburgers, to determine whether an increase in advertising expenditures will increase profit.
3. The estimated intercept implies that if both price and advertising expenditure were zero the sales revenue would be \$118,914. Clearly, this outcome is not possible; a zero price implies zero sales revenue. In this model, as in many others, it is important to recognize that the model is an approximation to reality in the region for which we have data. Including an intercept improves this approximation even when it is not directly interpretable.

In giving the above interpretations, we had to be careful to recognize the units of measurement for each of the variables. What would happen if we measured *PRICE* in cents instead of dollars and *SALES* in dollars instead of thousands of dollars? To discover the outcome, define the new variables measured in terms of the new units as $PRICE^* = 100 \times PRICE$ and $SALES^* = 1000 \times SALES$. Substituting for *PRICE* and *SALES*, our new fitted equation becomes

$$\frac{\widehat{SALES}^*}{1000} = 118.91 - 7.908\frac{PRICE^*}{100} + 1.863ADVERT$$

Multiplying through by 1000, we obtain

$$\widehat{SALES}^* = 118{,}910 - 79.08PRICE^* + 1863ADVERT$$

This is the estimated model that we would obtain if we applied least squares to the variables expressed in terms of the new units of measurement. The standard errors would change in the same way, but the R^2 will stay the same. In this form, a more direct interpretation of the coefficients is possible. A one cent increase in *PRICE* leads to a decline in mean *SALES* of \$79.08. An increase in *ADVERT* of \$1000 leads to an increase in mean sales revenue of \$1863.

TABLE 5.2 Least Squares Estimates for Sales Equation for Big Andy's Burger Barn

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	118.9136	6.3516	18.7217	0.0000
PRICE	−7.9079	1.0960	−7.2152	0.0000
ADVERT	1.8626	0.6832	2.7263	0.0080
$R^2 = 0.4483$	$SSE = 1718.943$	$\hat{\sigma} = 4.8861$	$s_y = 6.48854$	

In addition to providing information about how sales change when price or advertising change, the estimated equation can be used for prediction. Suppose Big Andy is interested in predicting sales revenue for a price of \$5.50 and an advertising expenditure of \$1200. Including extra decimal places to get an accurate hand calculation, this prediction is

$$\begin{aligned} SALES &= 118.91 - 7.908PRICE + 1.863ADVERT \\ &= 118.914 - 7.9079 \times 5.5 + 1.8626 \times 1.2 \\ &= 77.656 \end{aligned}$$

The predicted value of sales revenue for $PRICE = 5.5$ and $ADVERT = 1.2$ is \$77,656.

Remark

A word of caution is in order about interpreting regression results: The negative sign attached to price implies that reducing the price will increase sales revenue. If taken literally, why should we not keep reducing the price to zero? Obviously that would not keep increasing total revenue. This makes the following important point: Estimated regression models describe the relationship between the economic variables for values similar to those found in the sample data. Extrapolating the results to extreme values is generally not a good idea. Predicting the value of the dependent variable for values of the explanatory variables far from the sample values invites disaster. Refer to Figure 4.2 and the surrounding discussion.

5.2.2 Estimating the Error Variance σ^2

There is one remaining parameter to estimate—the variance of the error term. For this parameter, we follow the same steps that were outlined in Section 2.7. Under assumptions MR1, MR2, and MR3, we know that

$$\sigma^2 = \text{var}(e_i|\mathbf{X}) = \text{var}(e_i) = E(e_i^2|\mathbf{X}) = E(e_i^2)$$

Thus, we can think of σ^2 as the expectation or population mean of the squared errors e_i^2. A natural estimator of this population mean is the sample mean $\hat{\sigma}^2 = \sum e_i^2/N$. However, the squared errors e_i^2 are unobservable, so we develop an estimator for σ^2 that is based on their counterpart, the squares of the least squares residuals. For the model in (5.6), these residuals are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (b_1 + b_2 x_{i2} + b_3 x_{i3})$$

An estimator for σ^2 that uses the information from $\hat{e}_i^2$ and has good statistical properties is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N - K} \tag{5.11}$$

where K is the number of β parameters being estimated in the multiple regression model. We can think of $\hat{\sigma}^2$ as an average of $\hat{e}_i^2$ with the denominator in the averaging process being $N - K$ instead of N. It can be shown that replacing e_i^2 by $\hat{e}_i^2$ requires the use of $N - K$ instead of N for $\hat{\sigma}^2$ to be unbiased. Note that in equation (2.19), where there was only one explanatory variable and two coefficients, we had $K = 2$.

To appreciate further why $\hat{e}_i$ provide information about σ^2, recall that σ^2 measures the variation in e_i or, equivalently, the variation in y_i around the mean function $\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$. Since $\hat{e}_i$ are estimates of e_i, big values of $\hat{e}_i$ suggest σ^2 is large while small $\hat{e}_i$ suggest σ^2 is small. When we refer to "big" values of $\hat{e}_i$, we mean big positive ones or big negative ones. Using the squares of the residuals $\hat{e}_i^2$ means that positive values do not cancel with negative ones; thus, $\hat{e}_i^2$ provide information about the parameter σ^2.

EXAMPLE 5.3 | Error Variance Estimate for Hamburger Chain Data

In the hamburger chain example, we have $K = 3$. The estimate for our sample of data in Table 5.1 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N - K} = \frac{1718.943}{75 - 3} = 23.874$$

Go back and have a look at Table 5.2. There are two quantities in this table that relate to the above calculation. The first is the **sum of squared errors**

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2 = 1718.943$$

The second is the square root of $\hat{\sigma}^2$, given by

$$\hat{\sigma} = \sqrt{23.874} = 4.8861$$

Both these quantities typically appear in the output from your computer software. Different software refer to it in different ways. Sometimes $\hat{\sigma}$ is referred to as the **standard error of the regression**. Sometimes it is called the **root mse** (short for the square root of mean squared error).

5.2.3 Measuring Goodness-of-Fit

For the simple regression model studied in Chapter 4, we introduced R^2 as a measure of the proportion of variation in the dependent variable that is explained by variation in the explanatory variable. In the multiple regression model the same measure is relevant, and the same formulas are valid, but now we talk of the proportion of variation in the dependent variable explained by *all* the explanatory variables included in the model. The coefficient of determination is

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}\hat{e}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5.12}$$

where SSR is the variation in y "explained" by the model (**sum of squares due to regression**), SST is the total variation in y about its mean (sum of squares, total), and SSE is the sum of squared least squares residuals (errors) and is that part of the variation in y that is not explained by the model.

The notation $\hat{y}_i$ refers to the predicted value of y for each of the sample values of the explanatory variables, that is,

$$\hat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_K x_{iK}$$

The sample mean $\bar{y}$ is both the mean of the y_i and the mean of the $\hat{y}_i$, providing the model that includes an intercept $(\beta_1$ in this case$)$.

The value for SSE will be reported by almost all computer software, but sometimes SST is not reported. Recall, however, that the sample standard deviation for y, which is readily computed by most software, is given by

$$s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} = \sqrt{\frac{SST}{N-1}}$$

and so

$$SST = (N-1)s_y^2$$

EXAMPLE 5.4 | R^2 for Hamburger Chain Data

Using the results for Big Andy's Burger Barn in Table 5.2, we find that $SST = 74 \times 6.48854^2 = 3115.485$ and $SSE = 1718.943$. Using these sums of squares, we have

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{e}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{1718.943}{3115.485} = 0.448$$

The interpretation of R^2 is that 44.8% of the variation in sales revenue about its mean is explained by the variation in price and the variation in the level of advertising expenditure. It means that, *in our sample*, 55.2% of the variation in revenue is left unexplained and is due to variation in the error term or variation in other variables that implicitly form part of the error term.

As mentioned in Section 4.2.2, the coefficient of determination is also viewed as a measure of the predictive ability of the model over the sample period, or as a measure of how well the estimated regression fits the data. The value of R^2 is equal to the squared sample correlation coefficient between $\hat{y}_i$ and y_i. Since the sample correlation measures the linear association between two variables, if the R^2 is high, that means there is a close association between the values of y_i and the values predicted by the model, $\hat{y}_i$. In this case, the model is said to "fit" the data well. If R^2 is low, there is not a close association between the values of y_i and the values predicted by the model, $\hat{y}_i$, and the model does not fit the data well.

One final note is in order. The intercept parameter β_1 is the y-intercept of the regression "plane," as shown in Figure 5.1. If, for theoretical reasons, you are *certain* that the regression plane passes through the origin, then $\beta_1 = 0$ and it can be omitted from the model. While this is not a common practice, it does occur, and regression software includes an option that removes the intercept from the model. If the model does not contain an intercept parameter, then the measure R^2 given in (5.12) is no longer appropriate. The reason it is no longer appropriate is that, without an intercept term in the model,

$$\sum_{i=1}^{N} (y_i - \bar{y})^2 \neq \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N} \hat{e}_i^2$$

or, $SST \neq SSR + SSE$. To understand why, go back and check the proof in Appendix 4B. In the sum of squares decomposition the cross-product term $\sum_{i=1}^{N} (\hat{y}_i - \bar{y})\hat{e}_i$ no longer disappears. Under these circumstances, it does not make sense to talk of the proportion of total variation that is explained by the regression. Thus, when your model does not contain a constant, it is better not to report R^2, even if your computer displays one.

5.2.4 Frisch–Waugh–Lovell (FWL) Theorem

The Frisch–Waugh–Lovell (FWL) Theorem[5] is a useful and somewhat surprising result that we use a number of times in the remainder of the book. It also helps understand in a multiple regression the interpretation of a coefficient estimate, *all other variables held constant*. To illustrate[6]

[5] Also known as the Frisch–Waugh Theorem or the decomposition theorem.

[6] An illustration is not a proof. For a nonmatrix algebra proof, see Michael C. Lovell (2008) "A Simple Proof of the FWL Theorem," *Journal of Economic Education*, Winter 2008, 88–91. A proof using matrix algebra is presented in William H. Greene (2018) *Econometric Analysis, Eighth Edition*, Boston: Prentice-Hall, 36–38.

this result, we use the sales equation $SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i$ and carry out the following steps:

1. Estimate the simple regression $SALES_i = a_1 + a_2 PRICE_i + error$ using the least squares estimator and save the least squares residuals.
$$\widetilde{SALES}_i = SALES_i - (\hat{a}_1 + \hat{a}_2 PRICE_i) = SALES_i - (121.9002 - 7.8291 PRICE_i)$$
2. Estimate the simple regression $ADVERT_i = c_1 + c_2 PRICE_i + error$ using the least squares estimator and save the least squares residuals.
$$\widetilde{ADVERT}_i = ADVERT_i - (\hat{c}_1 + \hat{c}_2 PRICE_i) = ADVERT_i - (1.6035 + 0.0423 PRICE_i)$$
3. Estimate the simple regression $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$ with no constant term. The estimate of β_3 is $b_3 = 1.8626$. This estimate is the same as that reported from the full regression in Table 5.2.
4. Compute the least squares residuals from step 3, $\hat{\tilde{e}}_i = \widetilde{SALES}_i - b_3 \widetilde{ADVERT}_i$. Compare these residuals to those from the complete model.
$$\hat{e}_i = SALES_i - (b_1 + b_2 PRICE_i + b_3 ADVERT_i)$$

You will find that the two sets of residuals $\hat{\tilde{e}}_i$ and $\hat{e}_i$ are identical. Consequently, the sums of squared residuals are also the same, $\sum \hat{e}_i^2 = \sum \hat{\tilde{e}}_i^2 = 1718.943$.

What have we shown?

- In steps 1 and 2, we removed (or "purged" or "partialled out") the linear influence of *PRICE* (and a constant term) from both *SALES* and *ADVERT* by estimating least squares regressions and computing the least squares residuals $\widetilde{SALES}$ and $\widetilde{ADVERT}$. These residual variables are *SALES* and *ADVERT* after removing, or "partialling out," the linear influence of *PRICE* and a constant.
- In step 3, we illustrate the first important result of the **FWL theorem**: the coefficient estimate for β_3 from the regression using the partialled-out variables $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$ is exactly the same as that from the full regression $SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i$. We have explained β_3 as "the change in monthly sales *SALES* ($1000) when advertising expenditure *ADVERT* is increased by one unit ($1000), and the price index *PRICE* **is held constant**." The FWL result gives a precise meaning to "is held constant." It means that β_3 is the effect of advertising expenditure on sales after the linear influence of price and a constant term have been removed from both.
- In step 4, we note the second important result of the FWL theorem: the least squares residuals and their sum of squares are identical when calculated from the full regression or the "partialled-out" model.

A few cautions are in order. First, pay attention to the constant term. Here we have included it with *PRICE* as a variable to be partialled out in steps 1 and 2. Consequently, a constant is not included in step 3. Second, estimating the partialled-out regression is not *completely* equivalent to estimating the original, complete model. When estimating $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$, your software will see only one parameter to estimate, β_3. Consequently, when computing the estimate of σ^2, software will use the degrees of freedom $N - 1 = 74$. This means that the reported estimated error variance will be too small. It is $\tilde{\sigma}^2 = \sum \hat{\tilde{e}}_i^2/(N-1) = 1718.943/74 = 23.2290$ compared to the estimate from the previous section that uses divisor $N - K = 75 - 3$, $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-3) = 1718.943/72 = 23.8742$.[7] Third, for illustration we have used estimates that are rounded to four decimals. In practice, your software will use more significant digits. The results of the theorem may suffer from rounding error if insufficient significant digits are used. The estimate in step 3 is

[7] This smaller error variance estimate means that the standard errors of the regression coefficients discussed in Section 5.3.1 will be too small.

accurate to four decimals in this example, but the least squares residuals in step 4 are off without using more significant digits.

The Frisch–Waugh–Lovell Theorem also applies in the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$. Partition the explanatory variables into two groups. The theorem works for any partition, but generally the variables that are not the primary focus of the analysis are partialled out. This group is sometimes called the collection of **control variables** as they are included for a proper specification of the regression model and "control for" the variables that are not of primary interest. For example, suppose that x_2 and x_3 are the variables of primary interest. Then the two groups are $g_1 = (x_{i2}, x_{i3})$ and $g_2 = (x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK})$. Note that we have included the constant term in group two but not group one. Each variable must go into one group or the other but not both. The FWL theorem is then applied in the following steps:

1. Estimate the least squares regression with dependent variable y and the explanatory variables $g_2 = (x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK})$. Compute the least squares residuals, $\tilde{y}$.
2. Estimate the least squares regression for each variable in group one using explanatory variables $g_2 = (x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK})$ and compute the least squares residuals, $\tilde{x}_2$ and $\tilde{x}_3$.
3. Estimate the least squares regression using the partialled-out variables, $\tilde{y}_i = \beta_2 \tilde{x}_{i2} + \beta_3 \tilde{x}_{i3} + \tilde{e}_i$. The coefficient estimates b_2 and b_3 will be identical to the estimates from the full model.
4. The residuals from the partialled-out regression, $\hat{\tilde{e}}_i = \tilde{y}_i - (b_2 \tilde{x}_{i2} + b_3 \tilde{x}_{i3})$, are identical to the residuals from the full model.

5.3 Finite Sample Properties of the Least Squares Estimator

In a general context, the least squares estimators (b_1, b_2, b_3) are random variables; they take on different values in different samples and their values are unknown until a sample is collected and their values computed. The differences from sample to sample are called "sampling variation" and are unavoidable. The probability or **sampling distribution** of the OLS estimator describes how its estimates vary over all possible samples. The **sampling properties** of the OLS estimator refer to characteristics of this distribution. If the mean of the distribution of b_k is β_k, the estimator is unbiased. The variance of the distribution provides a basis for assessing the reliability of the estimates. If the variability of b_k across samples is relatively high, then it is hard to be confident that the values obtained in one realized sample will necessarily be close to the true parameters. On the other hand, if b_k is unbiased and its variability across samples is relatively low, we can be confident that an estimate from one sample will be reliable.

What we can say about the sampling distribution of the least squares estimator depends on what assumptions can realistically be made for the sample of data being used for estimation. For the simple regression model introduced in Chapter 2 we saw that, under the assumptions SR1 to SR5, the OLS estimator is best linear unbiased in the sense that there is no other linear unbiased estimator that has a lower variance. The same result holds for the general multiple regression model under assumptions MR1–MR5.

The Gauss–Markov Theorem: If assumptions MR1–MR5 hold, the least squares estimators are the **B**est **L**inear **U**nbiased **E**stimators (BLUE) of the parameters in the multiple regression model.[8]

[8] Similar remarks can be made about the properties of the least squares estimator in the multiple regression model under the more restrictive, but sometimes realistic, assumptions explored for the simple regression model in Section 2.10. Under the assumptions in that section, if all explanatory variables are statistically independent of all error terms, or if the observations on $(y_i, x_{i2}, x_{i3}, \ldots, x_{iK})$ are collected via random sampling making them independent, the BLUE property still holds.

The implications of adding assumption MR6, that the errors are normally distributed, are also similar to those from the corresponding assumption made for the simple regression model. Conditional on $\mathbf{X}$, the least squares estimator is normally distributed. Using this result, and the **error variance estimator** $\hat{\sigma}^2$, a t-statistic that follows a t-distribution can be constructed and used for interval estimation and hypothesis testing, along similar lines to the development in Chapter 3.

These various properties—BLUE and the use of the t-distribution for interval estimation and hypothesis testing—are **finite sample** properties. As long as $N > K$, they hold irrespective of the sample size N. We provide more details in the context of the multiple regression model in the remainder of this section and in Sections 5.4 and 5.5. There are, however, many circumstances where we are unable to rely on finite sample properties. Violation of some of the assumptions can mean that finite sample properties of the OLS estimator do not hold or are too difficult to derive. Also, as we travel through the book and encounter more complex models and assumptions designed for a variety of different types of sample data, an ability to use finite sample properties becomes the exception rather than the rule. To accommodate such situations we use what are called **large sample** or **asymptotic properties**. These properties refer to the behavior of the sampling distribution of an estimator as the sample size approaches infinity. Under less restrictive assumptions, or when faced with a more complex model, large sample properties can be easier to derive than finite sample properties. Of course, we never have infinite samples, but the idea is that if N is sufficiently large, then an estimator's properties as N becomes infinite will be a good approximation to that estimator's properties when N is large but finite. We discuss large sample properties and the circumstances under which they need to be invoked in Section 5.7. An example is the central limit theorem mentioned in Section 2.6. There we learnt that, if N is sufficiently large, the least squares estimator is approximately normally distributed even when assumption SR6, which specifies that the errors are normally distributed, is violated.

5.3.1 The Variances and Covariances of the Least Squares Estimators

The variances and covariances of the least squares estimators give us information about the reliability of the estimators b_1, b_2, and b_3. Since the least squares estimators are unbiased, the smaller their variances, the higher the probability that they will produce estimates "near" the true parameter values. For $K = 3$, we can express the conditional variances and covariances in an algebraic form that provides useful insights into the behavior of the least squares estimator. For example, we can show that

$$\operatorname{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{(1 - r_{23}^2)\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2} \tag{5.13}$$

where r_{23} is the sample correlation coefficient between the values of x_2 and x_3; see Section 4.2.1. Its formula is given by

$$r_{23} = \frac{\sum(x_{i2} - \bar{x}_2)(x_{i3} - \bar{x}_3)}{\sqrt{\sum(x_{i2} - \bar{x}_2)^2 \sum(x_{i3} - \bar{x}_3)^2}}$$

For the other variances and covariances, there are formulas of a similar nature. It is important to understand the factors affecting the variance of b_2:

1. Larger error variances σ^2 lead to larger variances of the least squares estimators. This is to be expected, since σ^2 measures the overall uncertainty in the model specification. If σ^2 is large, then data values may be widely spread about the regression function $E(y_i|\mathbf{X}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$, and there is less information in the data about the parameter values. If σ^2 is small, then data values are compactly spread about the regression function $E(y_i|\mathbf{X}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$, and there is more information about what the parameter values might be.

2. Larger sample sizes N imply smaller variances of the least squares estimators. A larger value of N means a larger value of the summation $\sum(x_{i2} - \bar{x}_2)^2$. Since this term appears in the denominator of (5.13), when it is large, $\text{var}(b_2)$ is small. This outcome is also an intuitive one; more observations yield more precise parameter estimation.
3. More variation in an explanatory variable around its mean, measured in this case by $\sum(x_{i2} - \bar{x}_2)^2$, leads to a smaller variance of the least squares estimator. To estimate β_2 precisely, we prefer a large amount of variation in x_{i2}. The intuition here is that if the variation or change in x_2 is small, it is difficult to measure the effect of that change. This difficulty will be reflected in a large variance for b_2.
4. A larger correlation between x_2 and x_3 leads to a larger variance of b_2. Note that $1 - r_{23}^2$ appears in the denominator of (5.13). A value of $|r_{23}|$ close to 1 means $1 - r_{23}^2$ will be small, which in turn means $\text{var}(b_2)$ will be large. The reason for this fact is that variation in x_{i2} about its mean adds most to the precision of estimation when it is not connected to variation in the other explanatory variables. When the variation in one explanatory variable is connected to variation in another explanatory variable, it is difficult to disentangle their separate effects. In Chapter 6, we discuss "collinearity," which is the situation when the explanatory variables are correlated with one another. Collinearity leads to increased variances of the least squares estimators.

Although our discussion has been in terms of a model where $K = 3$, these factors affect the variances of the least squares estimators in the same way in larger models.

It is customary to arrange the estimated variances and covariances of the least squares estimators in a square array, which is called a matrix. This matrix has variances on its diagonal and covariances in the off-diagonal positions. It is called a **variance–covariance matrix** or, more simply, a **covariance matrix**. When $K = 3$, the arrangement of the variances and covariances in the covariance matrix is

$$\text{cov}(b_1, b_2, b_3) = \begin{bmatrix} \text{var}(b_1) & \text{cov}(b_1, b_2) & \text{cov}(b_1, b_3) \\ \text{cov}(b_1, b_2) & \text{var}(b_2) & \text{cov}(b_2, b_3) \\ \text{cov}(b_1, b_3) & \text{cov}(b_2, b_3) & \text{var}(b_3) \end{bmatrix}$$

Before discussing estimation of this matrix, it is useful to distinguish between the covariance matrix conditional on the observed explanatory variables $\text{cov}(b_1, b_2, b_3|\mathbf{X})$, and the unconditional covariance matrix $\text{cov}(b_1, b_2, b_3)$ that recognizes that most data generation is such that both y and $\mathbf{X}$ are random variables. Given that the OLS estimator is both conditionally and unconditionally unbiased, that is, $E(b_k) = E(b_k|\mathbf{X}) = \beta_k$, the unconditional covariance matrix is given by

$$\text{cov}(b_1, b_2, b_3) = E_{\mathbf{X}}[\text{cov}(b_1, b_2, b_3|\mathbf{X})]$$

Taking the variance of b_2 as an example of one of the elements in this matrix, we have

$$\text{var}(b_2) = E_{\mathbf{X}}[\text{var}(b_2|\mathbf{X})] = \sigma^2 E_{\mathbf{X}}\left[\frac{1}{(1 - r_{23}^2)\sum_{i=1}^N (x_{i2} - \bar{x}_2)^2}\right]$$

We use the same quantity to estimate both $\text{var}(b_2)$ and $\text{var}(b_2|\mathbf{X})$. That is,

$$\widehat{\text{var}}(b_2) = \widehat{\text{var}}(b_2|\mathbf{X}) = \frac{\hat{\sigma}^2}{(1 - r_{23}^2)\sum_{i=1}^N (x_{i2} - \bar{x}_2)^2}$$

This quantity is an unbiased estimator for both $\text{var}(b_2)$ and $\text{var}(b_2|\mathbf{X})$. For estimating $\text{var}(b_2|\mathbf{X})$, we replace σ^2 with $\hat{\sigma}^2$ in equation (5.13). For estimating $\text{var}(b_2)$, we replace σ^2 with $\hat{\sigma}^2$ *and* the unknown expectation $E_{\mathbf{X}}\left\{\left[(1 - r_{23}^2)\sum_{i=1}^N (x_{i2} - \bar{x}_2)^2\right]^{-1}\right\}$ with $\left[(1 - r_{23}^2)\sum_{i=1}^N (x_{i2} - \bar{x}_2)^2\right]^{-1}$. Similar replacements are made for the other elements in the covariance matrix.

EXAMPLE 5.5 | Variances, Covariances, and Standard Errors for Hamburger Chain Data

Using the estimate $\hat{\sigma}^2 = 23.874$ and our computer software package, the estimated variances and covariances for b_1, b_2, b_3, in the Big Andy's Burger Barn example are

$$\widehat{\text{cov}}(b_1, b_2, b_3) = \begin{bmatrix} 40.343 & -6.795 & -0.7484 \\ -6.795 & 1.201 & -0.0197 \\ -0.7484 & -0.0197 & 0.4668 \end{bmatrix}$$

Thus, we have

$$\widehat{\text{var}}(b_1) = 40.343 \qquad \widehat{\text{cov}}(b_1, b_2) = -6.795$$
$$\widehat{\text{var}}(b_2) = 1.201 \qquad \widehat{\text{cov}}(b_1, b_3) = -0.7484$$
$$\widehat{\text{var}}(b_3) = 0.4668 \qquad \widehat{\text{cov}}(b_2, b_3) = -0.0197$$

Table 5.3 shows how this information is typically reported in the output from computer software. Of particular relevance are the standard errors of b_1, b_2, and b_3; they are given by the square roots of the corresponding estimated variances. That is,

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1)} = \sqrt{40.343} = 6.3516$$
$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2)} = \sqrt{1.201} = 1.0960$$
$$\text{se}(b_3) = \sqrt{\widehat{\text{var}}(b_3)} = \sqrt{0.4668} = 0.6832$$

Again, it is time to go back and look at Table 5.2. Notice that these values appear in the standard error column.

These standard errors can be used to say something about the range of the least squares estimates if we were to obtain more samples of 75 Burger Barns from different cities. For example, the standard error of b_2 is approximately

TABLE 5.3 Covariance Matrix for Coefficient Estimates

	C	*PRICE*	*ADVERT*
C	40.3433	−6.7951	−0.7484
PRICE	−6.7951	1.2012	−0.0197
ADVERT	−0.7484	−0.0197	0.4668

$\text{se}(b_2) = 1.1$. We know that the least squares estimator is unbiased, so its mean value is $E(b_2) = \beta_2$. Suppose b_2 is approximately normally distributed, then based on statistical theory we expect 95% of the estimates b_2, obtained by applying the least squares estimator to other samples, to be within approximately two standard deviations of the mean β_2. Given our sample, $2 \times \text{se}(b_2) = 2.2$, so we estimate that 95% of the b_2 values would lie within the interval $\beta_2 \pm 2.2$. It is in this sense that the estimated variance of b_2, or its corresponding standard error, tells us something about the reliability of the least squares estimates. If the difference between b_2 and β_2 can be large, b_2 is not reliable; if the difference between b_2 and β_2 is likely to be small, then b_2 is reliable. Whether a particular difference is "large" or "small" will depend on the context of the problem and the use to which the estimates are to be put. This issue is considered again in later sections when we use the estimated variances and covariances to test hypotheses about the parameters and to construct interval estimates.

5.3.2 The Distribution of the Least Squares Estimators

We have asserted that, under the multiple regression model assumptions MR1–MR5, listed in Section 5.1, the least squares estimator b_k is the best linear unbiased estimator of the parameter β_k in the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$$

If we add assumption MR6, that the random errors e_i have normal probability distributions, then, conditional on $\mathbf{X}$, the dependent variable y_i is normally distributed:

$$(y_i|\mathbf{X}) \sim N\left((\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}), \sigma^2\right) \iff (e_i|\mathbf{X}) \sim N(0, \sigma^2)$$

For a given $\mathbf{X}$, the least squares estimators are linear functions of dependent variables, which means that the conditional distribution of the least squares estimators is also normal:

$$(b_k|\mathbf{X}) \sim N\left(\beta_k, \text{var}(b_k|\mathbf{X})\right)$$

That is, given $\mathbf{X}$, each b_k has a normal distribution with mean β_k and variance $\text{var}(b_k|\mathbf{X})$. By subtracting its mean and dividing by the square root of its variance, we can transform the normal

random variable b_k into a standard normal variable Z with mean zero and a variance of one.

$$Z = \frac{b_k - \beta_k}{\sqrt{\text{var}(b_k|\mathbf{X})}} \sim N(0,1), \qquad \text{for } k = 1, 2, \ldots, K \tag{5.14}$$

What is particularly helpful about this result is that the distribution of Z does not depend on any unknown parameters or on $\mathbf{X}$. Although the unconditional distribution of b_k will almost certainly not be normal—it depends on the distributions of both e and $\mathbf{X}$—we can use the standard normal distribution to make probability statements about Z irrespective of whether the explanatory variables are treated as fixed or random. As mentioned in Chapter 3, statistics with this property are called **pivotal**.

There is one remaining problem, however. Before we can use (5.14) to construct interval estimates for β_k or test hypothesized values for β_k, we need to replace the unknown parameter σ^2 that is a component of $\text{var}(b_k|\mathbf{X})$ with its estimator $\hat{\sigma}^2$. Doing so yields a t random variable given by

$$t = \frac{b_k - \beta_k}{\sqrt{\widehat{\text{var}}(b_k|\mathbf{X})}} = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-K)} \tag{5.15}$$

Like Z in equation (5.14), the distribution of this t-statistic does not depend on any unknown parameters or on $\mathbf{X}$. It is a generalization of the result in equation (3.2). A difference is the degrees of freedom of the t random variable. In Chapter 3, where there were two coefficients to be estimated, the number of degrees of freedom was $(N-2)$. In this chapter, there are K unknown coefficients in the general model and *the number of degrees of freedom for t-statistics is* $(N-K)$.

Linear Combinations of Parameters The result in (5.15) extends to a linear combination of coefficients that was introduced in Section 3.6. Suppose that we are interested in estimating or testing hypotheses about a linear combination of coefficients that in the general case is given by

$$\lambda = c_1\beta_1 + c_2\beta_2 + \cdots + c_K\beta_K = \sum_{k=1}^{K} c_k\beta_k$$

Then

$$t = \frac{\hat{\lambda} - \lambda}{\text{se}(\hat{\lambda})} = \frac{\sum c_k b_k - \sum c_k \beta_k}{\text{se}(\sum c_k b_k)} \sim t_{(N-K)} \tag{5.16}$$

This expression is a little intimidating, mainly because we have included all coefficients to make it general, and because hand calculation of $\text{se}(\sum c_k b_k)$ is onerous if more than two coefficients are involved. For example, if $K = 3$, then

$$\text{se}(c_1b_1 + c_2b_2 + c_3b_3) = \sqrt{\widehat{\text{var}}(c_1b_1 + c_2b_2 + c_3b_3|\mathbf{X})}$$

where

$$\begin{aligned}\widehat{\text{var}}(c_1b_1 + c_2b_2 + c_3b_3|\mathbf{X}) = {} & c_1^2\widehat{\text{var}}(b_1|\mathbf{X}) + c_2^2\widehat{\text{var}}(b_2|\mathbf{X}) + c_3^2\widehat{\text{var}}(b_3|\mathbf{X}) + 2c_1c_2\widehat{\text{cov}}(b_1, b_2|\mathbf{X}) \\ & + 2c_1c_3\widehat{\text{cov}}(b_1, b_3|\mathbf{X}) + 2c_2c_3\widehat{\text{cov}}(b_2, b_3|\mathbf{X})\end{aligned}$$

In many instances some of the c_k will be zero, which can simplify the expressions and the calculations considerably. If one c_k is equal to one, and the rest are zero, (5.16) simplifies to (5.15).

What happens if the errors are not normally distributed? Then the least squares estimator will not be normally distributed and (5.14), (5.15), and (5.16) will not hold exactly. They will, however, be approximately true in large samples. Thus, having errors that are not normally distributed does not stop us from using (5.15) and (5.16), but it does mean we have to be cautious if the sample size is not large. A test for normally distributed errors was given in Section 4.3.5. An example of errors that are not normally distributed can be found in Appendix 5C.

We now examine how the results in (5.15) and (5.16) can be used for interval estimation and hypothesis testing. The procedures are identical to those described in Chapter 3, except that the degrees of freedom change.

5.4 Interval Estimation

5.4.1 Interval Estimation for a Single Coefficient

Suppose we are interested in finding a 95% **interval estimate** for β_2, the response of average sales revenue to a change in price at Big Andy's Burger Barn. Following the procedures described in Section 3.1, and noting that we have $N - K = 75 - 3 = 72$ degrees of freedom, the first step is to find a value from the $t_{(72)}$-distribution, call it t_c, such that

$$P(-t_c < t_{(72)} < t_c) = 0.95 \tag{5.17}$$

Using the notation introduced in Section 3.1, $t_c = t_{(0.975,\, N-K)}$ is the 97.5-percentile of the $t_{(N-K)}$-distribution (the area or probability to the left of t_c is 0.975), and $-t_c = t_{(0.025,\, N-K)}$ is the 2.5-percentile of the $t_{(N-K)}$-distribution (the area or probability to the left of $-t_c$ is 0.025). Consulting the t-table (Statistical Table 2), we discover there is no entry for 72 degrees of freedom, but, from the entries for 70 and 80 degrees of freedom, it is clear that, correct to two decimal places, $t_c = 1.99$. If greater accuracy is required, your computer software can be used to find $t_c = 1.993$. Using this value, and the result in (5.15) for the second coefficient ($k = 2$), we can rewrite (5.17) as

$$P\left(-1.993 \le \frac{b_2 - \beta_2}{\text{se}(b_2)} \le 1.993\right) = 0.95$$

Rearranging this expression, we obtain

$$P\left[b_2 - 1.993 \times \text{se}(b_2) \le \beta_2 \le b_2 + 1.993 \times \text{se}(b_2)\right] = 0.95$$

The interval endpoints

$$\left[b_2 - 1.993 \times \text{se}(b_2),\ \ b_2 + 1.993 \times \text{se}(b_2)\right] \tag{5.18}$$

define a 95% interval estimator of β_2. If this interval estimator is used in many samples from the population, then 95% of them will contain the true parameter β_2. We can establish this fact before any data are collected, based on the model assumptions alone. Before the data are collected, we have confidence in the **interval estimation procedure (estimator)** because of its performance over all possible samples.

EXAMPLE 5.6 | Interval Estimates for Coefficients in Hamburger Sales Equation

A 95% interval estimate for β_2 based on our particular sample is obtained from (5.18) by replacing b_2 and $\text{se}(b_2)$ by their values $b_2 = -7.908$ and $\text{se}(b_2) = 1.096$. Thus, our 95% interval estimate for β_2 is given by[9]

$$(-7.9079 - 1.9335 \times 1.096,\ 7.9079 + 1.9335 \times 1.096)$$
$$= (-10.093, -5.723)$$

This interval estimate suggests that decreasing price by \$1 will lead to an increase in average revenue somewhere between \$5723 and \$10,093. Or, in terms of a price change whose magnitude is more realistic, a 10-cent price reduction will lead to an average revenue increase between \$572 and \$1009. Based on this information, and the cost of making and selling more burgers, Big Andy can decide whether to proceed with a price reduction.

[9]For this and the next calculation, we used more digits so that it would match the more accurate computer output. You may see us do this occasionally.

Following a similar procedure for β_3, the response of average sales revenue to advertising, we find a 95% interval estimate is given by

$$(1.8626 - 1.9935 \times 0.6832,\ 1.8626 + 1.9935 \times 0.6832)$$
$$= (0.501,\ 3.225)$$

We estimate that an increase in advertising expenditure of \$1000 leads to an increase in average sales revenue of between \$501 and \$3225. This interval is a relatively wide one; it implies that extra advertising expenditure could be unprofitable (the revenue increase is less than \$1000) or could lead to a revenue increase more than three times the cost of the advertising. Another way of describing this situation is to say that the point estimate $b_3 = 1.8626$ is not very reliable, as its standard error (which measures sampling variability) is relatively large.

In general, if an interval estimate is uninformative because it is too wide, there is nothing immediate that can be done. A wide interval for the parameter β_3 arises because the estimated sampling variability of the least squares estimator b_3 is large. In the computation of an interval estimate, a large sampling variability is reflected by a large standard error. A narrower interval can only be obtained by reducing the variance of the estimator. Based on the variance expression in (5.13), one solution is to obtain more and better data exhibiting more independent variation. Big Andy could collect data from other cities and set a wider range of price and advertising combinations. It might be expensive to do so, however, and so he would need to assess whether the extra information is worth the extra cost. This solution is generally not open to economists, who rarely use controlled experiments to obtain data. Alternatively, we might introduce some kind of nonsample information on the coefficients. The question of how to use both sample and nonsample information in the estimation process is taken up in Chapter 6.

We cannot say, in general, what constitutes an interval that is too wide, or too uninformative. It depends on the context of the problem being investigated, and on how the information is to be used.

To give a general expression for an interval estimate, we need to recognize that the **critical value** t_c will depend on the degree of confidence specified for the interval estimate and the number of degrees of freedom. We denote the degree of confidence by $1 - \alpha$; in the case of a 95% interval estimate $\alpha = 0.05$ and $1 - \alpha = 0.95$. The number of degrees of freedom is $N - K$; in Big Andy's Burger Barn example this value was $75 - 3 = 72$. The value t_c is the percentile value $t_{(1-\alpha/2,\, N-K)}$, which has the property that $P\left[t_{(N-K)} \le t_{(1-\alpha/2,\, N-K)}\right] = 1 - \alpha/2$. In the case of a 95% confidence interval, $1 - \alpha/2 = 0.975$; we use this value because we require 0.025 in each tail of the distribution. Thus, we write the general expression for a $100(1-\alpha)\%$ confidence interval as

$$\left[b_k - t_{(1-\alpha/2,N-K)} \times \mathrm{se}(b_k),\ b_k + t_{(1-\alpha/2,\, N-K)} \times \mathrm{se}(b_k)\right]$$

5.4.2 Interval Estimation for a Linear Combination of Coefficients

The t-statistic in (5.16) can also be used to create interval estimates for a variety of linear combinations of parameters. Such combinations are of interest if we are considering the value of $E(y|\mathbf{X})$ for a particular setting of the explanatory variables, or the effect of changing two or more explanatory variables simultaneously. They become especially relevant if the effect of an explanatory variable depends on two or more parameters, a characteristic of many nonlinear relationships that we explore in Section 5.6.

EXAMPLE 5.7 | Interval Estimate for a Change in Sales

Big Andy wants to make next week a big sales week. He plans to increase advertising expenditure by \$800 and drop the price by 40 cents. If the prices before and after the changes are $PRICE_0$ and $PRICE_1$, respectively, and those for advertising expenditure are $ADVERT_0$ and $ADVERT_1$, then the change in expected sales from Andy's planned strategy is

$$\begin{aligned}\lambda &= E(SALES_1|PRICE_1, ADVERT_1) \\ &\quad - E(SALES_0|PRICE_0, ADVERT_0) \\ &= [\beta_1 + \beta_2 PRICE_1 + \beta_3 ADVERT_1] \\ &\quad - [\beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0] \\ &= [\beta_1 + \beta_2 (PRICE_0 - 0.4) + \beta_3 (ADVERT_0 + 0.8)] \\ &\quad - [\beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0] \\ &= -0.4\beta_2 + 0.8\beta_3\end{aligned}$$

Andy would like a point estimate and a 90% interval estimate for λ.

A point estimate is given by

$$\begin{aligned}\hat{\lambda} &= -0.4b_2 + 0.8b_3 = -0.4 \times (-7.9079) + 0.8 \times 1.8626 \\ &= 4.6532\end{aligned}$$

Our estimate of the expected increase in sales from Big Andy's strategy is \$4653.

From (5.16), we can derive a 90% interval estimate for $\lambda = -0.4\beta_2 + 0.8\beta_3$ as

$$\begin{aligned}&\left[\hat{\lambda} - t_c \times \text{se}(\hat{\lambda}),\ \hat{\lambda} + t_c \times \text{se}(\hat{\lambda})\right] \\ &\quad = \left[(-0.4b_2 + 0.8b_3) - t_c \times \text{se}(-0.4b_2 + 0.8b_3),\right. \\ &\qquad \left.(-0.4b_2 + 0.8b_3) + t_c \times \text{se}(-0.4b_2 + 0.8b_3)\right]\end{aligned}$$

where $t_c = t_{(0.95,\,72)} = 1.666$. Using the covariance matrix of the coefficient estimates in Table 5.3, and the result for the variance of a linear function of two random variables—see equation (3.8)—we can calculate the standard error $\text{se}(-0.4b_2 + 0.8b_3)$ as follows:

$$\begin{aligned}&\text{se}(-0.4b_2 + 0.8b_3) \\ &\quad = \sqrt{\widehat{\text{var}}(-0.4b_2 + 0.8b_3|\mathbf{X})} \\ &\quad = \left[(-0.4)^2 \widehat{\text{var}}(b_2|\mathbf{X}) + (0.8)^2 \widehat{\text{var}}(b_3|\mathbf{X})\right. \\ &\qquad \left. -2 \times 0.4 \times 0.8 \times \widehat{\text{cov}}(b_2, b_3|\mathbf{X})\right]^{1/2} \\ &\quad = [0.16 \times 1.2012 + 0.64 \times 0.4668 - 0.64 \times (-0.0197)]^{1/2} \\ &\quad = 0.7096\end{aligned}$$

Thus, a 90% interval estimate is

$$\begin{aligned}&(4.6532 - 1.666 \times 0.7096,\ 4.6532 + 1.666 \times 0.7096) \\ &\quad = (3.471,\ 5.835)\end{aligned}$$

We estimate, with 90% confidence, that the expected increase in sales from Big Andy's strategy will lie between \$3471 and \$5835.

5.5 Hypothesis Testing

As well as being useful for interval estimation, the t-distribution result in (5.15) provides the foundation for testing hypotheses about individual coefficients. As you discovered in Chapter 3, hypotheses of the form $H_0: \beta_2 = c$ versus $H_1: \beta_2 \neq c$, where c is a specified constant, are called two-tail tests. Hypotheses with inequalities such as $H_0: \beta_2 \leq c$ versus $H_1: \beta_2 > c$ are called one-tail tests. In this section, we consider examples of each type of hypothesis. For a **two-tail test**, we consider testing the significance of an individual coefficient; for one-tail tests, some hypotheses of economic interest are considered. Using the result in (5.16), one- and two-tail tests can also be used to test hypotheses about linear combinations of coefficients. An example of this type follows those for testing hypotheses about individual coefficients. We will follow the step-by-step procedure for testing hypotheses that was introduced in Section 3.4. To refresh your memory, here are the steps again:

Step-by-Step Procedure for Testing Hypotheses

1. Determine the null and alternative hypotheses.
2. Specify the test statistic and its distribution if the null hypothesis is true.
3. Select α and determine the rejection region.
4. Calculate the sample value of the test statistic and, if desired, the p-value.
5. State your conclusion.

At the time these steps were introduced, in Chapter 3, you had not discovered p-values. Knowing about p-values (see Section 3.5) means that steps 3–5 can be framed in terms of the test statistic and its value and/or the p-value. We will use both.

5.5.1 Testing the Significance of a Single Coefficient

When we set up a multiple regression model, we do so because we believe that the explanatory variables influence the dependent variable y. If we are to confirm this belief, we need to examine whether or not it is supported by the data. That is, we need to ask whether the data provide any evidence to suggest that y is related to each of the explanatory variables. If a given explanatory variable, say x_k, has no bearing on y, then $\beta_k = 0$. Testing this null hypothesis is sometimes called a test of significance for the explanatory variable x_k. Thus, to find whether the data contain any evidence suggesting y is related to x_k, we test the null hypothesis

$$H_0 : \beta_k = 0$$

against the alternative hypothesis

$$H_1 : \beta_k \neq 0$$

To carry out the test, we use the test statistic (5.15), which, if the null hypothesis is true, is

$$t = \frac{b_k}{\text{se}(b_k)} \sim t_{(N-K)}$$

For the alternative hypothesis "not equal to," we use a two-tail test, introduced in Section 3.3.3, and reject H_0 if the computed t-value is greater than or equal to t_c (the critical value from the right side of the distribution) or less than or equal to $-t_c$ (the critical value from the left side of the distribution). For a test with level of significance α, $t_c = t_{(1-\alpha/2,\, N-K)}$ and $-t_c = t_{(\alpha/2,\, N-K)}$. Alternatively, if we state the acceptance–rejection rule in terms of the p-value, we reject H_0 if $p \leq \alpha$ and do not reject H_0 if $p > \alpha$.

EXAMPLE 5.8 | Testing the Significance of Price

In the Big Andy's Burger Barn example, we test, following our standard testing format, whether sales revenue is related to price:

1. The null and alternative hypotheses are $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 \neq 0$.
2. The test statistic, if the null hypothesis is true, is $t = b_2/\text{se}(b_2) \sim t_{(N-K)}$.
3. Using a 5% significance level ($\alpha = 0.05$), and noting that there are 72 degrees of freedom, the critical values that lead to a probability of 0.025 in each tail of the distribution are $t_{(0.975,\, 72)} = 1.993$ and $t_{(0.025,\, 72)} = -1.993$. Thus, we reject the null hypothesis if the calculated value of t from step 2 is such that $t \geq 1.993$ or $t \leq -1.993$. If $-1.993 < t < 1.993$, we do not reject H_0. Stating the acceptance–rejection rule in terms of the p-value, we reject H_0 if $p \leq 0.05$ and do not reject H_0 if $p > 0.05$.
4. The computed value of the t-statistic is

$$t = \frac{-7.908}{1.096} = -7.215$$

From your computer software, the p-value in this case can be found as

$$P\left(t_{(72)} > 7.215\right) + P\left(t_{(72)} < -7.215\right) = 2 \times \left(2.2 \times 10^{-10}\right) = 0.000$$

Correct to three decimal places the result is p-value $= 0.000$.

5. Since $-7.215 < -1.993$, we reject $H_0 : \beta_2 = 0$ and conclude that there is evidence from the data to suggest that sales revenue depends on price. Using the p-value to perform the test, we reject H_0 because $0.000 < 0.05$.

EXAMPLE 5.9 | Testing the Significance of Advertising Expenditure

For testing whether sales revenue is related to advertising expenditure, we have

1. $H_0 : \beta_3 = 0$ and $H_1 : \beta_3 \neq 0$.
2. The test statistic, if the null hypothesis is true, is $t = b_3 / \text{se}(b_3) \sim t_{(N-K)}$.
3. Using a 5% significance level, we reject the null hypothesis if $t \geq 1.993$ or $t \leq -1.993$. In terms of the p-value, we reject H_0 if $p \leq 0.05$. Otherwise, we do not reject H_0.
4. The value of the test statistic is

$$t = \frac{1.8626}{0.6832} = 2.726$$

The p-value is given by

$$P(t_{(72)} > 2.726) + P(t_{(72)} < -2.726) = 2 \times 0.004 = 0.008$$

5. Because $2.726 > 1.993$, we reject H_0; the data support the conjecture that revenue is related to advertising expenditure. The same test outcome can be obtained using the p-value. In this case, we reject H_0 because $0.008 < 0.05$.

Note that the t-values -7.215 (Example 5.8) and 2.726 and their corresponding p-values 0.000 and 0.008 were reported in Table 5.2 at the same time that we reported the original least squares estimates and their standard errors. Hypothesis tests of this kind are carried out routinely by computer software, and their outcomes can be read immediately from the computer output that will be similar to Table 5.2.

When we reject a hypothesis of the form $H_0 : \beta_k = 0$, we say that the estimate b_k is significant. Significance of a coefficient estimate is desirable—it confirms an initial prior belief that a particular explanatory variable is a relevant variable to include in the model. However, we cannot be absolutely certain that $\beta_k \neq 0$. There is still a probability α that we have rejected a true null hypothesis. Also, as mentioned in Section 3.4, statistical significance of an estimated coefficient should not be confused with the economic importance of the corresponding explanatory variable. If the estimated response of sales revenue to advertising had been $b_3 = 0.01$ with a standard error of $\text{se}(b_3) = 0.005$, then we would have concluded that b_3 is significantly different from zero; but, since the estimate implies increasing advertising by \$1000 increases revenue by only \$10, we would not conclude that advertising is important. We should also be cautious about concluding that statistical significance implies precise estimation. The advertising coefficient $b_3 = 1.8626$ was found to be significantly different from zero, but we also concluded that the corresponding 95% interval estimate (0.501, 3224) was too wide to be very informative. In other words, we were not able to get a precise estimate of β_3.

5.5.2 One-Tail Hypothesis Testing for a Single Coefficient

In Section 5.1, we noted that two important considerations for the management of Big Andy's Burger Barn were whether demand was price-elastic or price-inelastic and whether the additional sales revenue from additional advertising expenditure would cover the costs of the advertising. We are now in a position to state these questions as testable hypotheses, and to ask whether the hypotheses are compatible with the data.

EXAMPLE 5.10 | Testing for Elastic Demand

With respect to demand elasticity, we wish to know whether

- $\beta_2 \geq 0$: a decrease in price leads to a change in sales revenue that is zero or negative (demand is price-inelastic or has an elasticity of unity).
- $\beta_2 < 0$: a decrease in price leads to an increase in sales revenue (demand is price-elastic).

The fast food industry is very competitive with many substitutes for Andy's burgers. We anticipate elastic demand and

put this conjecture as the alternative hypothesis. Following our standard testing format, we first state the null and alternative hypotheses:

1. $H_0: \beta_2 \geq 0$ (demand is unit-elastic or inelastic)
   $H_1: \beta_2 < 0$ (demand is elastic)
2. To create a test statistic, we act as if the null hypothesis is the equality $\beta_2 = 0$. Doing so is valid because if we reject H_0 for $\beta_2 = 0$, we also reject it for any $\beta_2 > 0$. Then, assuming that $H_0: \beta_2 = 0$ is true, from (5.15) the test statistic is $t = b_2/\text{se}(b_2) \sim t_{(N-K)}$.
3. The rejection region consists of values from the t-distribution that are unlikely to occur if the null hypothesis is true. If we define "unlikely" in terms of a 5% significance level, then unlikely values of t are those less than the critical value $t_{(0.05,\,72)} = -1.666$. Thus, we reject H_0 if $t \leq -1.666$ or if the p-value ≤ 0.05.
4. The value of the test statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{-7.908}{1.096} = -7.215$$

   The corresponding p-value is $P(t_{(72)} < -7.215) = 0.000$.
5. Since $-7.215 < -1.666$, we reject $H_0: \beta_2 \geq 0$ and conclude that $H_1: \beta_2 < 0$ (demand is elastic) is more compatible with the data. The sample evidence supports the proposition that a reduction in price will bring about an increase in sales revenue. Since $0.000 < 0.05$, the same conclusion is reached using the p-value.

Note the similarities and differences between this test and the two-tail test of significance performed in Section 5.5.1. The calculated t-values are the same, but the critical t-values are different. Not only are the values themselves different, but with a two-tail test there are also two critical values, one from each side of the distribution. With a one-tail test there is only one critical value, from one side of the distribution. Also, the p-value from the one-tail test is usually, but not always, half that of the two-tail test, although this fact is hard to appreciate from this example because both p-values are essentially zero.

EXAMPLE 5.11 | Testing Advertising Effectiveness

The other hypothesis of interest is whether an increase in advertising expenditure will bring an increase in sales revenue that is sufficient to cover the increased cost of advertising. We want proof that our advertising is profitable. If not, we may change advertising firms. Since advertising will be profitable if $\beta_3 > 1$, we set up the hypotheses:

1. $H_0: \beta_3 \leq 1$ and $H_1: \beta_3 > 1$.
2. Treating the null hypothesis as the equality $H_0: \beta_3 = 1$, the test statistic that has the t-distribution when H_0 is true is, from (5.15),

$$t = \frac{b_3 - 1}{\text{se}(b_3)} \sim t_{(N-K)}$$

3. Choosing $\alpha = 0.05$ as our level of significance, the relevant critical value is $t_{(0.95,\,72)} = 1.666$. We reject H_0 if $t \geq 1.666$ or if the p-value ≤ 0.05.
4. The value of the test statistic is

$$t = \frac{b_3 - \beta_3}{\text{se}(b_3)} = \frac{1.8626 - 1}{0.6832} = 1.263$$

   The p-value of the test is $P(t_{(72)} > 1.263) = 0.105$.
5. Since $1.263 < 1.666$, we do not reject H_0. There is insufficient evidence in our sample to conclude that advertising will be cost-effective. Using the p-value to perform the test, we again conclude that H_0 cannot be rejected, because $0.105 > 0.05$. Another way of thinking about the test outcome is as follows: Because the estimate $b_3 = 1.8626$ is greater than one, this estimate by itself suggests that advertising will be effective. However, when we take into account the precision of estimation, measured by the standard error, we find that $b_3 = 1.8626$ is not significantly greater than one. In the context of our hypothesis-testing framework, we cannot conclude with a sufficient degree of certainty that $\beta_3 > 1$.

5.5.3 Hypothesis Testing for a Linear Combination of Coefficients

We are often interested in testing hypotheses about linear combinations of coefficients. Will particular settings of the explanatory variables lead to a mean value of the dependent variable above

a certain threshold? Will changes in the values of two or more explanatory variables lead to a mean dependent variable change that exceeds a predefined goal? The t-statistic in (5.16) can be used to answer these questions.

EXAMPLE 5.12 | Testing the Effect of Changes in Price and Advertising

Big Andy's marketing adviser claims that dropping the price by 20 cents will be more effective for increasing sales revenue than increasing advertising expenditure by \$500. In other words, she claims that $-0.2\beta_2 > 0.5\beta_3$. Andy does not wish to accept this proposition unless it can be verified by past data. He knows that the estimated change in expected sales from the price fall is $-0.2b_2 = -0.2 \times (-7.9079) = 1.5816$, and that the estimated change in expected sales from the extra advertising is $0.5b_3 = 0.5 \times 1.8626 = 0.9319$, so the marketer's claim appears to be correct. However, he wants to establish whether the difference $1.5816 - 0.9319$ could be attributable to sampling error, or whether it constitutes proof, at a 5% significance level, that $-0.2\beta_2 > 0.5\beta_3$. This constitutes a test about a linear combination of coefficients. Since $-0.2\beta_2 > 0.5\beta_3$ can be written as $-0.2\beta_2 - 0.5\beta_3 > 0$, we are testing a hypothesis about the linear combination $-0.2\beta_2 - 0.5\beta_3$.

Following our hypothesis testing steps, we have

1. $H_0: -0.2\beta_2 - 0.5\beta_3 \leq 0$ (the marketer's claim is not correct)
   $H_1: -0.2\beta_2 - 0.5\beta_3 > 0$ (the marketer's claim is correct)
2. Using (5.16) with $c_2 = -0.2$, $c_3 = -0.5$ and all other c_k's equal to zero, and assuming that the equality in H_0 holds $(-0.2\beta_2 - 0.5\beta_3 = 0)$, the test statistic and its distribution when H_0 is true are

$$t = \frac{-0.2b_2 - 0.5b_3}{\text{se}(-0.2b_2 - 0.5b_3)} \sim t_{(72)}$$

3. For a one-tail test and a 5% significance level, the critical value is $t_{(0.95,\,72)} = 1.666$. We reject H_0 if $t \geq 1.666$ or if the p-value ≤ 0.05.
4. To find the value of the test statistic, we first compute

$$\begin{aligned}
&\text{se}(-0.2b_2 - 0.5b_3) \\
&= \sqrt{\widehat{\text{var}}(-0.2b_2 - 0.5b_3|\mathbf{X})} \\
&= \left[(-0.2)^2\widehat{\text{var}}(b_2|\mathbf{X}) + (-0.5)^2\widehat{\text{var}}(b_3|\mathbf{X})\right. \\
&\quad \left. +2 \times (-0.2) \times (-0.5) \times \widehat{\text{cov}}(b_2, b_3|\mathbf{X})\right]^{1/2} \\
&= \left[0.04 \times 1.2012 + 0.25 \times 0.4668 + 0.2 \times (-0.0197)\right]^{1/2} \\
&= 0.4010
\end{aligned}$$

   Then, the value of the test statistic is

$$t = \frac{-0.2b_2 - 0.5b_3}{\text{se}(-0.2b_2 - 0.5b_3)} = \frac{1.58158 - 0.9319}{0.4010} = 1.622$$

   The corresponding p-value is $P(t_{(72)} > 1.622) = 0.055$.
5. Since $1.622 < 1.666$, we do not reject H_0. At a 5% significance level, there is not enough evidence to support the marketer's claim. Alternatively, we reach the same conclusion using the p-value, because $0.055 > 0.05$.

5.6 Nonlinear Relationships

The multiple regression model that we have studied so far has the form

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e \tag{5.19}$$

It is a linear function of variables (the x's) and of the coefficients (the β's) and e. However, (5.19) is much more flexible than it at first appears. Although the assumptions of the multiple regression model require us to retain the property of linearity in the β's, many different nonlinear functions of variables can be specified by defining the x's and/or y as transformations of original variables. Several examples of such transformations have already been encountered for the simple regression model. In Chapter 2, the quadratic model $y = \alpha_1 + \alpha_2 x^2 + e$ and the log-linear model $\ln(y) = \gamma_1 + \gamma_2 x + e$ were estimated. A detailed analysis of these and other nonlinear simple regression models—a linear-log model, a log-log model, and a cubic model—was given in Chapter 4. The same kind of variable transformations and interpretations of their coefficients carry over to multiple regression models. One class of models is that of **polynomial** equations

such as the quadratic $y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$ or the cubic $y = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \alpha_4 x^3 + e$. When we studied these models as examples of the simple regression model, we were constrained by the need to have only one right-hand-side variable, such as $y = \beta_1 + \beta_3 x^2 + e$ or $y = \alpha_1 + \alpha_4 x^3 + e$. Now that we are working within the framework of the multiple regression model, we can consider unconstrained polynomials with all their terms included. Another generalization is to include "cross-product" or "interaction" terms leading to a model such as $y = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_2 x_3 + e$. In this section, we explore a few of the many options that are available for modeling nonlinear relationships. We begin with some examples of polynomial functions from economics. Polynomials are a rich class of functions that can parsimoniously describe relationships that are curved, with one or more peaks and valleys.

EXAMPLE 5.13 | Cost and Product Curves

In microeconomics, you studied "cost" curves and "product" curves that describe a firm. Total cost and total product curves are mirror images of each other, taking the standard "cubic" shapes shown in Figure 5.2. Average and marginal cost curves, and their mirror images, average and marginal product curves, take quadratic shapes, usually represented as shown in Figure 5.3.

The slopes of these relationships are not constant and cannot be represented by regression models that are "linear in the variables." However, these shapes are easily represented by polynomials. For example, if we consider the average cost relationship in Figure 5.3(a), a suitable regression model is

$$AC = \beta_1 + \beta_2 Q + \beta_3 Q^2 + e \tag{5.20}$$

This quadratic function can take the "U" shape we associate with average cost functions. For the total cost curve in Figure 5.2(a), a cubic polynomial is in order,

$$TC = \alpha_1 + \alpha_2 Q + \alpha_3 Q^2 + \alpha_4 Q^3 + e \tag{5.21}$$

These functional forms, which represent nonlinear shapes, can still be estimated using the least squares methods we have

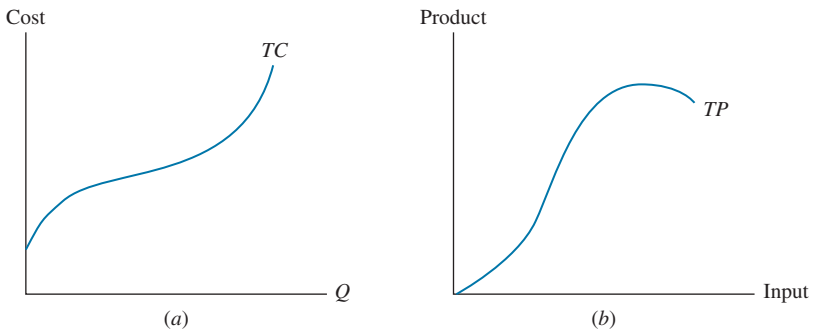


FIGURE 5.2 **(*a*) Total cost curve and (*b*) total product curve.**

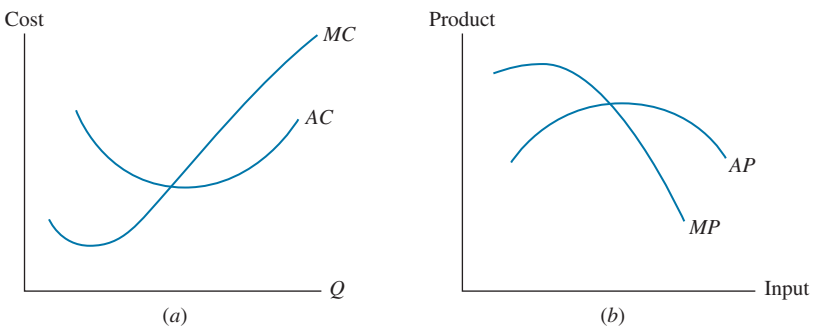


FIGURE 5.3 **Average and marginal (*a*) cost curves and (*b*) product curves.**

studied. The variables Q^2 and Q^3 are explanatory variables that are treated no differently from any others.

A difference in models of nonlinear relationships is in the interpretation of the parameters, which are not themselves slopes. To investigate the slopes, and to interpret the parameters, we need a little calculus. For the general polynomial function,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_p x^p$$

the slope or derivative of the curve is

$$\frac{dy}{dx} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + p a_p x^{p-1} \qquad (5.22)$$

This slope changes depending on the value of x. Evaluated at a particular value, $x = x_0$, the slope is

$$\left.\frac{dy}{dx}\right|_{x=x_0} = a_1 + 2a_2 x_0 + 3a_3 x_0^2 + \cdots + p a_p x_0^{p-1}$$

For more on rules of derivatives, see Appendix A.3.1.

Using the general rule in (5.22), the slope of the average cost curve (5.20) is

$$\frac{dE(AC)}{dQ} = \beta_2 + 2\beta_3 Q$$

The slope of the average cost curve changes for every value of Q and depends on the parameters β_2 and β_3. For this U-shaped curve, we expect $\beta_2 < 0$ and $\beta_3 > 0$. The slope of the total cost curve (5.21), which is the marginal cost, is

$$\frac{dE(TC)}{dQ} = \alpha_2 + 2\alpha_3 Q + 3\alpha_4 Q^2$$

The slope is a quadratic function of Q, involving the parameters α_2, α_3, and α_4. For a U-shaped marginal cost curve, we expect the parameter signs to be $\alpha_2 > 0$, $\alpha_3 < 0$, and $\alpha_4 > 0$.

Using polynomial terms is an easy and flexible way to capture nonlinear relationships between variables. As we have shown, care must be taken when interpreting the parameters of models that contain polynomial terms. Their inclusion does not complicate **least squares estimation**—with one exception. It is sometimes true that having a variable and its square or cube in the same model causes **collinearity** problems. (See Section 6.4.)

EXAMPLE 5.14 | Extending the Model for Burger Barn Sales

In the Burger Barn model $SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$, it is worth questioning whether the *linear* relationship between sales revenue, price, and advertising expenditure is a good approximation of reality. Having a linear model implies that increasing advertising expenditure will continue to increase sales revenue at the same rate irrespective of the existing levels of sales revenue and advertising expenditure—that is, that the coefficient β_3, which measures the response of $E(SALES|PRICE, ADVERT)$ to a change in *ADVERT*, is constant; it does not depend on the level of *ADVERT*. In reality, as the level of advertising expenditure increases, we would expect diminishing returns to set in. To illustrate what is meant by diminishing returns, consider the relationship between sales and advertising (assuming a fixed price) graphed in Figure 5.4. The figure shows the effect on sales of an increase of $200 in advertising expenditure when the original level of advertising is (a) $600 and (b) $1,600. Note that the units in the graph are thousands

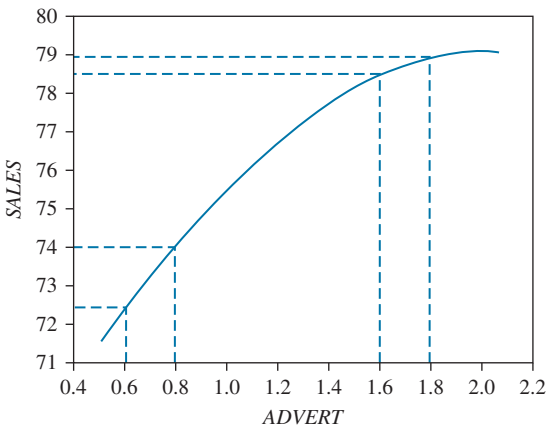


FIGURE 5.4 **A model where sales exhibits diminishing returns to advertising expenditure.**

of dollars, so these points appear as 0.6 and 1.6. At the smaller level of advertising, sales increase from \$72,400 to \$74,000, whereas at the higher level of advertising, the increase is a much smaller one, from \$78,500 to \$79,000. The linear model with the constant slope β_3 does not capture the diminishing returns.

What is required is a model where the slope changes as the level of *ADVERT* increases. One such model having this characteristic is obtained by including the squared value of advertising as another explanatory variable, making the new model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{5.23}$$

Adding the term $\beta_4 ADVERT^2$ to our original specification yields a model in which the response of expected revenue to a change in advertising expenditure depends on the level of advertising. Specifically, by applying the polynomial derivative rule in (5.22), and holding *PRICE* constant, the response of $E(SALES|PRICE, ADVERT)$ to a change in *ADVERT* is

$$\left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta ADVERT}\right|_{(PRICE \text{ held constant})} = \frac{\partial E(SALES|PRICE, ADVERT)}{\partial ADVERT} = \beta_3 + 2\beta_4 ADVERT \tag{5.24}$$

The partial derivative sign "∂" is used in place of the derivative sign "d" that we used in (5.22) because *SALES* depends on two variables, *PRICE* and *ADVERT*, and we are holding *PRICE* constant. See Appendix A.3.5 for further details about partial derivatives.

We refer to $\partial E(SALES|PRICE, ADVERT)/\partial ADVERT$ in (5.24) as the **marginal effect** of advertising on sales. In linear functions, the slope or marginal effect is constant. In nonlinear functions, it varies with one or more of the variables. To find the expected signs for β_3 and β_4, note that we expect the response of sales revenue to a change in advertising to be positive when $ADVERT = 0$. That is, we expect $\beta_3 > 0$. Also, to achieve diminishing returns, the response must decline as *ADVERT* increases. That is, we expect $\beta_4 < 0$.

Using least squares to estimate (5.23) yields

$$\underset{(se)}{\widehat{SALES}} = \underset{(6.80)}{109.72} - \underset{(1.046)}{7.640 PRICE} + \underset{(3.556)}{12.151 ADVERT} - \underset{(0.941)}{2.768 ADVERT^2} \tag{5.25}$$

What can we say about the addition of $ADVERT^2$ to the equation? Its coefficient has the expected negative sign and is significantly different from zero at a 5% significance level. Moreover, the coefficient of *ADVERT* has retained its positive sign and continues to be significant. The estimated response of sales to advertising is

$$\frac{\partial \widehat{SALES}}{\partial ADVERT} = 12.151 - 5.536 ADVERT$$

Substituting into this expression we find that when advertising is at its minimum value in the sample of \$500 ($ADVERT = 0.5$), the marginal effect of advertising on sales is 9.383. When advertising is at a level of \$2000 ($ADVERT = 2$), the marginal effect is 1.079. Thus, allowing for diminishing returns to advertising expenditure has improved our model both statistically and in terms of meeting our expectations about how sales will respond to changes in advertising.

EXAMPLE 5.15 | An Interaction Variable in a Wage Equation

In the last example, we saw how the inclusion of $ADVERT^2$ in the regression model for *SALES* has the effect of making the marginal effect of *ADVERT* on *SALES* depend on the level of *ADVERT*. What if the marginal effect of one variable depends on the level of another variable? How do we model it? To illustrate, consider a wage equation relating *WAGE* (\$ earnings per hour) to years of education (*EDUC*) and years of experience (*EXPER*) in the following way:

$$WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 (EDUC \times EXPER) + e \tag{5.26}$$

Here we are suggesting that the effect of another year's experience on wage may depend on a worker's level of education, and, similarly, the effect of another year of education may depend on the number of years of experience. Specifically,

$$\frac{\partial E(WAGE|EDUC, EXPER)}{\partial EXPER} = \beta_3 + \beta_4 EDUC$$

$$\frac{\partial E(WAGE|EDUC, EXPER)}{\partial EDUC} = \beta_2 + \beta_4 EXPER$$

Using the Current Population Survey data (*cps5_small*) to estimate (5.26), we obtain

$$\underset{(se)}{\widehat{WAGE}} = \underset{(4.162)}{-18.759} + \underset{(0.2833)}{2.6557 EDUC} + \underset{(0.1335)}{0.2384 EXPER} - \underset{(0.009400)}{0.002747 (EDUC \times EXPER)}$$

The negative estimate $b_4 = -0.002747$ suggests that the greater the number of years of education, the less valuable is an extra year of experience. Similarly, the greater the number of years of experience, the less valuable is an extra year of education. For a person with eight years of education, we estimate that an additional year of experience leads to an increase in average wages of $0.2384 - 0.002747 \times 8 = \$0.22$, whereas for a person with 16 years of education, the approximate increase in wages from an extra year of experience is $0.2384 - 0.002747 \times 16 = \$0.19$. For someone with no experience, the extra average wage from an extra year of education is \$2.66. The value of an extra year of education falls to $2.6557 - 0.002747 \times 20 = \$2.60$ for someone with 20 years of experience. These differences are not large. Perhaps there is no interaction effect—its estimated coefficient is not significantly different from zero—or perhaps we could improve the specification of the model.

EXAMPLE 5.16 | A Log-Quadratic Wage Equation

In equation (5.26), we used *WAGE* as the dependent variable whereas, when we previously studied a wage equation in Example 4.10, ln(*WAGE*) was chosen as the dependent variable. Labor economists tend to prefer ln(*WAGE*), believing that a change in years of education or experience is more likely to lead to a constant percentage change in *WAGE* than a constant absolute change. Also, a wage distribution will typically be heavily skewed to the right. Taking logarithms yields a distribution, which is shaped more like a normal distribution.

In the following example, we make two changes to the model in (5.26). We replace *WAGE* with ln(*WAGE*), and we add the variable $EXPER^2$. Adding $EXPER^2$ is designed to capture diminishing returns to extra years of experience. An extra year of experience for an old hand with many years of experience is likely to be less valuable than it would be for a rookie with limited or no experience. Thus, we specify the model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 (EDUC \times EXPER) + \beta_5 EXPER^2 + e \quad (5.27)$$

Here the two marginal effects which, when multiplied by 100 give the approximate percentage changes in wages from extra years of experience and education, respectively, are

$$\frac{\partial E[\ln(WAGE)\,|EDUC, EXPER]}{\partial EXPER} = \beta_3 + \beta_4 EDUC + 2\beta_5 EXPER \quad (5.28)$$

$$\frac{\partial E[\ln(WAGE)\,|EDUC, EXPER]}{\partial EDUC} = \beta_2 + \beta_4 EXPER \quad (5.29)$$

Having both the interaction term and the square of *EXPER* in the equation means that the marginal effect for experience will depend on both the level of education and the number of years of experience. Estimating (5.27) using the data in *cps5_small* yields

$$\begin{aligned} \widehat{\ln(WAGE)} &= 0.6792 + 0.1359 EDUC + 0.04890 EXPER \\ \text{(se)} \quad & (0.1561) \quad (0.0101) \quad (0.00684) \\ & - 0.001268 (EDUC \times EXPER) \\ & (0.000342) \\ & - 0.0004741 EXPER^2 \\ & (0.0000760) \end{aligned}$$

In this case, all estimates are significantly different from zero. Estimates of the percentage changes in wages from extra years of experience and extra years of education, computed using (5.28) and (5.29) for $EDUC = 8$ and 16 and $EXPER = 0$ and 20, are presented in Table 5.4. As expected, the value of an extra year of experience is greatest for someone with 8 years of education and no experience and smallest for someone with 16 years of education and 20 years of experience. We estimate that the value of an extra year of education is $13.59 - 11.06 = 2.53$ percentage points less for someone with 20 years of experience relative to someone with no experience.

TABLE 5.4 Percentage Changes in Wage

		% ΔWAGE/ΔEXPER		%ΔWAGE/ΔEDUC
		Years of education		
		8	16	
Years of experience	0	3.88	2.86	13.59
	20	1.98	0.96	11.06

5.7 Large Sample Properties of the Least Squares Estimator

It is nice to be able to use the finite sample properties of the OLS estimator or, indeed, any other estimator, to make inferences about population parameters[10]. Provided our assumptions are correct, we can be confident that we are basing our conclusions on procedures that are exact, whatever the sample size. However, the assumptions we have considered so far are likely to be too restrictive for many data sets. To accommodate less restrictive assumptions, as well as carry out inference for general functions of parameters, we need to examine the properties of estimators as sample size approaches infinity. Properties as sample size approaches infinity provide a good guide to properties in large samples. They will always be an approximation, but it is an approximation that improves as sample size increases. Large sample approximate properties are known as **asymptotic properties**. A question students always ask and instructors always evade is "how large does the sample have to be?" Instructors are evasive because the answer depends on the model, the estimator, and the function of parameters that is of interest. Sometimes $N = 30$ is adequate; sometimes $N = 1000$ or larger could be necessary. Some illustrations are given in the Monte Carlo experiments in Appendix 5C. In Appendix 5D, we explain how bootstrapping can be used to check whether a sample size is large enough for asymptotic properties to hold.

In this section, we introduce some large sample (asymptotic) properties and then discuss some of the circumstances where they are necessary.

5.7.1 Consistency

When choosing econometric estimators, we do so with the objective in mind of obtaining an estimate that is close to the true but unknown parameter with high probability. Consider the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$. Suppose that for decision-making purposes we consider that obtaining an estimate of β_2 within "epsilon" of the true value is satisfactory. The probability of obtaining an estimate "close" to β_2 is

$$P\left(\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon\right) \tag{5.30}$$

An estimator is said to be **consistent** if this probability converges to 1 as the sample size $N \to \infty$. Or, using the concept of a limit, the estimator b_2 is consistent if

$$\lim_{N\to\infty} P\left(\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon\right) = 1 \tag{5.31}$$

What does this mean? In Figure 5.5, we depict the probability density functions $f\left(b_{N_i}\right)$ for the least squares estimator b_2 based on samples sizes $N_4 > N_3 > N_2 > N_1$. As the sample size increases the probability density function (*pdf*) becomes narrower. Why is that so? First of all, the least squares estimator is unbiased if MR1–MR5 hold, so that $E\left(b_2\right) = \beta_2$. This property is true in samples of all sizes. As the sample size changes, the center of the *pdf*s remains at β_2. However, as the sample size N gets larger, the variance of the estimator b_2 becomes smaller. The center of the *pdf* remains fixed at $E\left(b_2\right) = \beta_2$, and the variance decreases, resulting in probability density functions like $f\left(b_{N_i}\right)$. The probability that b_2 falls in the interval $\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon$ is the area under the *pdf* between these limits. As the sample size increases, the probability of b_2 falling within the limits increases toward 1. In large samples, we can say that the least squares estimator will provide an estimate close to the true parameter with high probability.

[10]This section contains advanced materials.

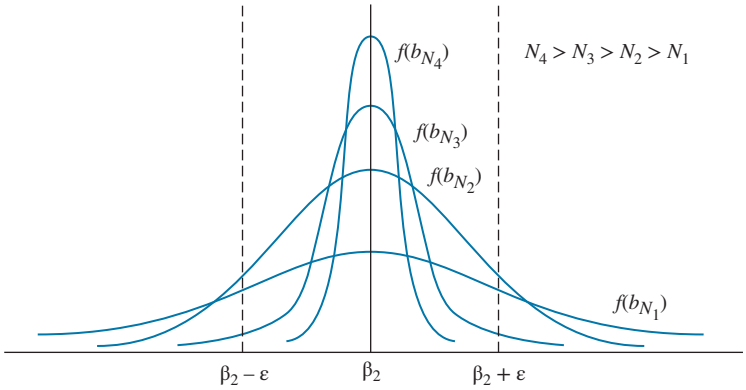


FIGURE 5.5 **An illustration of consistency.**

To appreciate why the variance decreases as N increases, consider the variance of the OLS estimator that we rewrite as follows:

$$\text{var}(b_2) = \sigma^2 E\left(\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right) = \frac{\sigma^2}{N} E\left(\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2/N}\right) = \frac{\sigma^2}{N} E\left[\left(s_x^2\right)^{-1}\right] = \frac{\sigma^2}{N} C_x \quad (5.32)$$

Notice that the N's that we have introduced cancel out. This trick is used so that we can write the variance for b_2 in terms of the sample variance of x, $s_x^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/N$.[11] Then, because $E\left[\left(s_x^2\right)^{-1}\right]$ is cumbersome, and a little intimidating, in the last equality we define the constant C_x as the expectation of the inverse of the sample variance. That is, $C_x = E\left[\left(s_x^2\right)^{-1}\right]$. The last result in (5.32) implies $\text{var}(b_2) \to 0$ as $N \to \infty$.

The property of **consistency** applies to many estimators, even ones that are biased in finite samples. For example, the estimator $\hat{\beta}_2 = b_2 + 1/N$ is a biased estimator. The amount of the bias is

$$\text{bias}\left(\hat{\beta}_2\right) = E\left(\hat{\beta}_2\right) - \beta_2 = \frac{1}{N}$$

For the estimator $\hat{\beta}_2$ the bias converges to zero as $N \to \infty$. That is,

$$\lim_{N\to\infty} \text{bias}\left(\hat{\beta}_2\right) = \lim_{N\to\infty}\left[E\left(\hat{\beta}_2\right) - \beta_2\right] = 0 \quad (5.33)$$

In this case, the estimator is said to be **asymptotically unbiased**. Consistency for an estimator can be established by showing that the estimator is either unbiased or asymptotically unbiased, and that its variance converges to zero as $N \to \infty$,

$$\lim_{N\to\infty} \text{var}\left(\hat{\beta}_2\right) = 0 \quad (5.34)$$

Conditions (5.33) and (5.34) are intuitive, and sufficient to establish an estimator to be consistent.

Because the probability density function of a consistent estimator collapses around the true parameter, and the probability that an estimator b_2 will be close to the true parameter β_2 approaches 1, the estimator b_2 is said to "converge in probability" to β_2, with the "in probability" part reminding us that it is the probability of being "close" in (5.31) that is the key factor. Several notations are used for this type of convergence. One is $b_2 \xrightarrow{p} \beta_2$, with the p over the arrow

[11] We have used N rather than $N - 1$ as the divisor for the sample variance. When dealing with properties as $N \to \infty$, it makes no difference which is used.

indicating "probability." A second is $\underset{N\to\infty}{\text{plim}}(b_2) = \beta_2$, with "plim" being short for "probability limit." Consistency is not just a large-sample alternative to unbiasedness; it is an important property in its own right. It is possible to find estimators that are unbiased but not consistent. The lack of consistency is considered undesirable even if an estimator is unbiased.

5.7.2 Asymptotic Normality

We mentioned earlier that the normal distribution assumption MR6: $(e_i|\mathbf{X}) \sim N(0, \sigma^2)$ is essential for the finite sample distribution of $(b_k|\mathbf{X})$ to be normal and for t-statistics such as $t = (b_k - \beta_k)/\text{se}(b_k)$ to have an exact t-distribution for use in interval estimation and hypothesis testing. However, we then went on to say that all is not lost if the normality assumption does not hold because, from a central limit theorem, the distribution of b_k will be approximately normal and interval estimates and t-tests will be approximately valid in large samples. Large sample approximate distributions are called **asymptotic distributions**. The need to use asymptotic distributions will become more urgent as we examine more complex models and estimators.

To appreciate how asymptotic distributions work and to introduce some notation, consider the OLS estimator b_2 in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$. We argued that the consistency of b_2 implies that the *pdf* for b_2 collapses to the point β_2 as $N \to \infty$. How can we get an approximate large sample distribution for b_2 if its *pdf* collapses to a single point? We consider instead the distribution of $\sqrt{N}b_2$. Note that $E(b_2) = \beta_2$ and that, from (5.32), $\text{var}(b_2) = \sigma^2 C_x/N$. It follows that $E\left(\sqrt{N}b_2\right) = \sqrt{N}\beta_2$ and

$$\text{var}\left(\sqrt{N}b_2\right) = \left(\sqrt{N}\right)^2 \text{var}(b_2) = N\sigma^2 C_x/N = \sigma^2 C_x$$

That is,

$$\sqrt{N}b_2 \sim \left(\sqrt{N}\beta_2, \sigma^2 C_x\right) \tag{5.35}$$

Central limit theorems are concerned with the distribution of sums (or averages) of random variables as $N \to \infty$.[12] In Chapter 2—see equation (2.12)—we showed that $b_2 = \beta_2 + \left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^{-1} \sum_{i=1}^N (x_i - \bar{x})\, e_i$ from which we can write

$$\sqrt{N}b_2 = \sqrt{N}\beta_2 + \left[s_x^2\right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N (x_i - \bar{x})\, e_i$$

Applying a central limit theorem to the sum $\sum_{i=1}^N (x_i - \bar{x})\, e_i/\sqrt{N}$, and using $\left[s_x^2\right]^{-1} \xrightarrow{p} C_x$, it can be shown that the statistic obtained by normalizing (5.35) so that it has mean zero and variance one, will be approximately normally distributed. Specifically,

$$\frac{\sqrt{N}(b_2 - \beta_2)}{\sqrt{\sigma^2 C_x}} \overset{a}{\sim} N(0, 1)$$

The notation $\overset{a}{\sim}$ is used to denote the asymptotic or approximate distribution. Recognizing that $\text{var}(b_2) = \sigma^2 C_x/N$, we can rewrite the above result as

$$\frac{(b_2 - \beta_2)}{\sqrt{\text{var}(b_2)}} \overset{a}{\sim} N(0, 1)$$

[12] There are several central limit theorems designed to accommodate sums of random variables with different properties. Their treatment is relatively advanced. See, for example, William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, online Appendix D.2.6, available at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

Going one step further, there is an important theorem that says replacing unknown quantities with consistent estimators does not change the asymptotic distribution of a statistic.[13] In this case, $\hat{\sigma}^2$ is a consistent estimator for σ^2 and $\left(s_x^2\right)^{-1}$ is a consistent estimator for C_x. Thus, we can write

$$t = \frac{\sqrt{N}(b_2 - \beta_2)}{\sqrt{\hat{\sigma}^2/s_x^2}} = \frac{(b_2 - \beta_2)}{\sqrt{\widehat{\text{var}}(b_2)}} = \frac{(b_2 - \beta_2)}{\text{se}(b_2)} \overset{a}{\sim} N(0, 1) \tag{5.36}$$

This is precisely the t-statistic that we use for interval estimation and hypothesis testing. The result in (5.36) means that using it in large samples is justified when assumption MR6 is not satisfied. One difference is that we are now saying that the distribution of the statistic "t" is approximately "normal," not "t." However, the t-distribution approaches the normal as $N \to \infty$, and it is customary to use either the t or the normal distribution as the large sample approximation. Because use of (5.36) for interval estimation or hypothesis testing implies we are behaving as if b_2 is normally distributed with mean β_2 and variance $\widehat{\text{var}}(b_2)$, this result is often written as

$$b_2 \overset{a}{\sim} N\left(\beta_2, \widehat{\text{var}}(b_2)\right) \tag{5.37}$$

Finally, our exposition has been in terms of the distribution of b_2 in the simple regression model, but the result also holds for estimators of the coefficients in the multiple regression model. In Appendix 5C, we use Monte Carlo experiments to illustrate how the central limit theorem works and give examples of how large N needs to be for the normal approximation to be satisfactory.

5.7.3 Relaxing Assumptions

In the previous two sections we explained that, when assumptions MR1–MR5 hold, and MR6 is relaxed, the least squares estimator is consistent and asymptotically normal. In this section, we investigate what we can say about the properties of the least squares estimator when we modify the strict exogeneity assumption MR2: $E(e_i|\mathbf{X}) = 0$ to make it less restrictive.

Weakening Strict Exogeneity: Cross-Sectional Data It is convenient to consider modifications of $E(e_i|\mathbf{X}) = 0$ first for cross-sectional data and then for time-series data. For cross-sectional data, we return to the random sampling assumptions, explained in Section 2.2, and written more formally in Section 2.10. Generalizing these assumptions to the multiple regression model, random sampling implies the joint observations $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ are independent, and that the strict exogeneity assumption $E(e_i|\mathbf{X}) = 0$ reduces to $E(e_i|\mathbf{x}_i) = 0$. Under this and the remaining assumptions of the model under random sampling, the least squares estimator is best linear unbiased. We now examine the implications of replacing $E(e_i|\mathbf{x}_i) = 0$ with the weaker assumption

$$E(e_i) = 0 \quad \text{and} \quad \text{cov}(e_i, x_{ik}) = 0 \quad \text{for } i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K \tag{5.38}$$

Why is (5.38) a weaker assumption? In Section 2.10, in the context of the simple regression model, we explained how $E(e_i|\mathbf{x}_i) = 0$ implies (5.38).[14] However, we cannot go back the other way. While $E(e_i|\mathbf{x}_i) = 0$ implies (5.38), (5.38) does not necessarily imply $E(e_i|\mathbf{x}_i) = 0$. Making the assumption $E(e_i|\mathbf{x}_i) = 0$ means that the best predictor for e_i is zero; there is no information in $\mathbf{x}_i$ that will help predict e_i. On the other hand, assuming $\text{cov}(e_i, x_{ik}) = 0$ only implies there is no *linear* predictor for e_i that is better than zero. It does not rule out nonlinear functions of $\mathbf{x}_i$ that may help predict e_i.

Why is it useful to consider the weaker assumption in (5.38)? First, the weaker are the assumptions under which an estimator has desirable properties, the wider the applicability of

[13] For more precise details, see William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorem D.16, in online Appendix available at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

[14] A proof is given in Appendix 2G.

the estimator. Second, as we discover in Chapter 10, violation of the assumption $\text{cov}(e_i, x_{ik}) = 0$ provides a good framework for considering the problem of endogenous regressors.

The seemingly innocuous weaker assumption in (5.38) means we can no longer show that the least squares estimator is unbiased. Consider the least squares estimator for β_2 in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. From (2.11) and (2.12),

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})e_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5.39}$$

and

$$E(b_2) = \beta_2 + E\left(\frac{\sum_{i=1}^{N}(x_i - \bar{x})e_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right) \tag{5.40}$$

Now, $E(e_i) = 0$ and $\text{cov}(e_i, x_{ik}) = 0$ imply $E(x_i e_i) = 0$, but the last term in (5.39) is more complicated than that; it involves the covariance between e_i and a function of x_i. This covariance will not necessarily be zero, implying $E(b_2) \neq \beta_2$. We can show that b_2 is consistent, however. We can rewrite (5.39) as

$$b_2 = \beta_2 + \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})e_i}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \beta_2 + \frac{\widehat{\text{cov}}(e_i, x_i)}{\widehat{\text{var}}(x_i)} \tag{5.41}$$

Because sample means, variances, and covariances computed from random samples are consistent estimators of their population counterparts,[15] we can say

$$\widehat{\text{cov}}(e_i, x_i) \xrightarrow{p} \text{cov}(e_i, x_i) = 0 \tag{5.42a}$$

$$\widehat{\text{var}}(x_i) \xrightarrow{p} \sigma_x^2 \tag{5.42b}$$

Thus, the second term in (5.41) converges in probability to zero, and $b_2 \xrightarrow{p} \beta_2$. It is also true that the asymptotic distribution of b_2 will be normal, as described in (5.36) and (5.37).

Weakening Strict Exogeneity: Time-Series Data When we turn to time-series data, the observations $(y_t, \mathbf{x}_t)$, $t = 1, 2, \ldots, T$ are not collected via random sampling and so it is no longer reasonable to assume they are independent. The explanatory variables will almost certainly be correlated over time, and the likelihood of the assumption $E(e_t|\mathbf{X}) = 0$ being violated is very strong indeed. To see why, note that $E(e_t|\mathbf{X}) = 0$ implies

$$E(e_t) = 0 \quad \text{and} \quad \text{cov}(e_t, x_{sk}) = 0 \quad \text{for} \quad t, s = 1, 2, \ldots, T; \quad k = 1, 2, \ldots, K \tag{5.43}$$

This result says that the errors in every time period are uncorrelated with all the explanatory variables in every time period. In Section 2.10.2, three examples of how this assumption might be violated were described. Now would be a good time to check out those examples. To reinforce them, consider the simple regression model $y_t = \beta_1 + \beta_2 x_t + e_t$, which is being estimated with time-series observations in periods $t = 1, 2, \ldots, T$. If x_t is a policy variable whose settings depend on past outcomes $y_{t-1}, y_{t-2}, \ldots$, then x_t will be correlated with previous errors $e_{t-1}, e_{t-2}, \ldots$. This is evident from the equation for the previous period observation $y_{t-1} = \beta_1 + \beta_2 x_{t-1} + e_{t-1}$. If x_t is correlated with y_{t-1}, then it will also be correlated with e_{t-1} since y_{t-1} depends directly on e_{t-1}. Such a correlation is particularly evident if x_t is a lagged value of y_t. That is, $y_t = \beta_1 + \beta_2 y_{t-1} + e_t$. Models of this type are called autoregressive models; they are considered in Chapter 9.

The likely violation of $\text{cov}(e_t, x_{sk}) = 0$ for $s \neq t$ implies $E(e_t|\mathbf{X}) = 0$ will be violated, which in turn implies we cannot show that the least squares estimator is unbiased. It is possible to show

[15] This result follows from a law of large numbers. See Theorem D.4 and its corollary in the online Appendix to William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, online at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm

it is consistent, however. To show consistency, we first assume that the errors and the explanatory variables in the same time period are uncorrelated. That is, we modify (5.43) to the less restrictive and more realistic assumption

$$E(e_t) = 0 \quad \text{and} \quad \text{cov}(e_t, x_{tk}) = 0 \quad \text{for} \quad t = 1, 2, \ldots, T; \; k = 1, 2, \ldots, K \tag{5.44}$$

Errors and the explanatory variables that satisfy (5.44) are said to be **contemporaneously uncorrelated**. We do not insist that $\text{cov}(e_t, x_{sk}) = 0$ for $t \neq s$. Now reconsider (5.41) written in terms of time-series observations

$$b_2 = \beta_2 + \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})e_t}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2} = \beta_2 + \frac{\widehat{\text{cov}}(e_t, x_t)}{\widehat{\text{var}}(x_t)} \tag{5.45}$$

Equation (5.45) is still valid, just as it was for cross-sectional observations. The question we need to ask to ensure consistency of b_2 is when the explanatory variables are not independent will it still be true that

$$\widehat{\text{cov}}(e_t, x_t) \xrightarrow{p} \text{cov}(e_t, x_t) = 0 \tag{5.46a}$$

$$\widehat{\text{var}}(x_t) \xrightarrow{p} \sigma_x^2 \tag{5.46b}$$

with σ_x^2 finite? The answer is "yes" as long as x is not "too dependent." If the correlation between the x_t's declines as they become further apart in time, then the results in (5.46) will hold. We reserve further discussion of the implications of the behavior of the explanatory variables in time-series regressions for Chapters 9 and 12. For the moment, we assume that their behavior is sufficiently cooperative for (5.46) to hold, so that the least squares estimator is consistent. At the same time, we recognize that, with time-series data, the least squares estimator is unlikely to be unbiased. **Asymptotic normality** can be shown by a central limit theorem, implying we can use (5.36) and (5.37) for interval estimation and hypothesis testing.

5.7.4 Inference for a Nonlinear Function of Coefficients

The need for large sample or asymptotic distributions is not confined to situations where assumptions MR1–MR6 are relaxed. Even if these assumptions hold, we still need to use large sample theory if a quantity of interest involves a nonlinear function of coefficients. To introduce this problem, we return to Big Andy's Burger Barn and examine the optimal level of advertising.

EXAMPLE 5.17 | The Optimal Level of Advertising

Economic theory tells us to undertake all those actions for which the marginal benefit is greater than the marginal cost. This optimizing principle applies to Big Andy's Burger Barn as it attempts to choose the optimal level of advertising expenditure. Recalling that *SALES* denotes sales revenue or total revenue, the marginal benefit in this case is the marginal revenue from more advertising. From (5.24), the required marginal revenue is given by the marginal effect of more advertising $\beta_3 + 2\beta_4 ADVERT$. The marginal cost of \$1 of advertising is \$1 plus the cost of preparing the additional products sold due to effective advertising. If we ignore the latter costs, the marginal cost of \$1 of advertising expenditure is \$1. Thus, advertising should be increased to the point where

$$\beta_3 + 2\beta_4 ADVERT_0 = 1$$

with $ADVERT_0$ denoting the optimal level of advertising. Using the least squares estimates for β_3 and β_4 in (5.25), a point estimate for $ADVERT_0$ is

$$\widehat{ADVERT_0} = \frac{1 - b_3}{2b_4} = \frac{1 - 12.1512}{2 \times (-2.76796)} = 2.014$$

implying that the optimal monthly advertising expenditure is \$2014.

To assess the reliability of this estimate, we need a standard error and an interval estimate for $(1-b_3)/2b_4$. This is a tricky problem, and one that requires the use of calculus to solve. What makes it more difficult than what we have done so far is the fact that it involves a **nonlinear function** of b_3 and b_4. Variances of nonlinear functions are hard to derive. Recall that the variance of a linear function, say, $c_3b_3 + c_4b_4$, is given by

$$\text{var}(c_3b_3 + c_4b_4) = c_3^2\text{var}(b_3) + c_4^2\text{var}(b_4) + 2c_3c_4\text{cov}(b_3, b_4) \tag{5.47}$$

Finding the variance of $(1-b_3)/2b_4$ is less straightforward. The best we can do is find an approximate expression that is valid in large samples. Suppose $\lambda = (1-\beta_3)/2\beta_4$ and $\hat{\lambda} = (1-b_3)/2b_4$; then, the approximate variance expression is

$$\text{var}(\hat{\lambda}) = \left(\frac{\partial\lambda}{\partial\beta_3}\right)^2\text{var}(b_3) + \left(\frac{\partial\lambda}{\partial\beta_4}\right)^2\text{var}(b_4) + 2\left(\frac{\partial\lambda}{\partial\beta_3}\right)\left(\frac{\partial\lambda}{\partial\beta_4}\right)\text{cov}(b_3, b_4) \tag{5.48}$$

This expression holds for all nonlinear functions of two estimators, not just $\hat{\lambda} = (1-b_3)/2b_4$. Also, note that for the linear case, where $\lambda = c_3\beta_3 + c_4\beta_4$ and $\hat{\lambda} = c_3b_3 + c_4b_4$, (5.48) reduces to (5.47). Using (5.48) to find an approximate expression for a variance is called the **delta method**. For further details, consult Appendix 5B.

We will use (5.48) to estimate the variance of $\hat{\lambda} = \widehat{ADVERT}_0 = (1-b_3)/2b_4$, get its standard error, and use that to get an interval estimate for $\lambda = ADVERT_0 = (1-\beta_3)/2\beta_4$. If the use of calculus in (5.48) frightens you, take comfort in the fact that most software will automatically compute the standard error for you.

The required derivatives are

$$\frac{\partial\lambda}{\partial\beta_3} = -\frac{1}{2\beta_4}, \quad \frac{\partial\lambda}{\partial\beta_4} = -\frac{1-\beta_3}{2\beta_4^2}$$

To estimate $\text{var}(\hat{\lambda})$, we evaluate these derivatives at the least squares estimates b_3 and b_4.

Thus, for the estimated variance of the optimal level of advertising, we have

$$\begin{aligned}\widehat{\text{var}}(\hat{\lambda}) &= \left(-\frac{1}{2b_4}\right)^2\widehat{\text{var}}(b_3) + \left(-\frac{1-b_3}{2b_4^2}\right)^2\widehat{\text{var}}(b_4) \\ &\quad + 2\left(-\frac{1}{2b_4}\right)\left(-\frac{1-b_3}{2b_4^2}\right)\widehat{\text{cov}}(b_3, b_4) \\ &= \left(\frac{1}{2\times 2.768}\right)^2 \times 12.646 \\ &\quad + \left(\frac{1-12.151}{2\times 2.768^2}\right)^2 \times 0.88477 \\ &\quad + 2\left(\frac{1}{2\times 2.768}\right)\left(\frac{1-12.151}{2\times 2.768^2}\right) \times 3.2887 \\ &= 0.016567\end{aligned}$$

and

$$\text{se}(\hat{\lambda}) = \sqrt{0.016567} = 0.1287$$

We are now in a position to get a 95% interval estimate for $\lambda = ADVERT_0$. When dealing with a linear combination of coefficients in (5.16), and Section 5.4.2, we used the result $(\hat{\lambda} - \lambda)/\text{se}(\hat{\lambda}) \sim t_{(N-K)}$. In line with Section 5.7.2, this result can be used in exactly the same way for nonlinear functions, but a difference is that the result is only an approximate one for large samples, even when the errors are normally distributed. Thus, an approximate 95% interval estimate for $ADVERT_0$ is

$$\begin{aligned}&\left[\hat{\lambda} - t_{(0.975,71)}\text{se}(\hat{\lambda}),\ \hat{\lambda} + t_{(0.975,71)}\text{se}(\hat{\lambda})\right] \\ &= [2.014 - 1.994\times 0.1287,\ 2.014 + 1.994\times 0.1287] \\ &= [1.757,\ 2.271]\end{aligned}$$

We estimate with 95% confidence that the optimal level of advertising lies between \$1757 and \$2271.

EXAMPLE 5.18 | How Much Experience Maximizes Wages?

In Example 5.16, we estimated the wage equation

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4(EDUC\times EXPER) + \beta_5 EXPER^2 + e$$

One of the implications of the quadratic function of experience is that, as a number of years of experience increases, wages will increase up to a point and then decline. Suppose we are interested in the number of years of experience, which maximizes *WAGE*. We can get this quantity by differentiating the wage equation with respect to *EXPER*, setting the first derivative equal to zero and solving for *EXPER*. It does not matter that the dependent variable is ln(*WAGE*) not *WAGE*; the value of *EXPER* that maximizes ln(*WAGE*) will also maximize *WAGE*. Setting the first derivative in (5.28) equal to zero and solving for *EXPER* yields

$$EXPER_0 = \frac{-\beta_3 - \beta_4 EDUC}{2\beta_5}$$

The maximizing value depends on the number of years of education. For someone with 16 years of education, it is

$$EXPER_0 = \frac{-\beta_3 - 16\beta_4}{2\beta_5}$$

Finding the standard error for an estimate of this function is tedious. It involves differentiating with respect to β_3, β_4, and β_5 and evaluating a variance expression involving three variances and three covariances—an extension of (5.48) to three coefficients. This is a problem better handled by your favorite econometric software. Taking this advice, we find $\widehat{EXPER_0} = 30.17$ and $\text{se}\left(\widehat{EXPER_0}\right) = 1.7896$. Then, a 95% interval estimate of the number of years of experience that maximizes *WAGE* is

$$\left[\widehat{EXPER_0} - t_{(0.975,1195)}\text{se}\left(\widehat{EXPER_0}\right),\ \widehat{EXPER_0} + t_{(0.975,1195)}\text{se}\left(\widehat{EXPER_0}\right)\right]$$

Inserting the relevant values yields

$$(30.17 - 1.962 \times 1.7896,\ 30.17 + 1.962 \times 1.7896)$$
$$= (26.7,\ 33.7)$$

We estimate that the number of years of experience that maximizes wages lies between 26.7 and 33.7 years.

5.8 Exercises

5.8.1 Problems

5.1 Consider the multiple regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + e_i$$

with the seven observations on y_i, x_{i1}, x_{i2}, and x_{i3} given in Table 5.5.

TABLE 5.5 Data for Exercise 5.1

y_i	x_{i1}	x_{i2}	x_{i3}
1	1	0	1
1	1	1	−2
4	1	2	2
0	1	−2	1
1	1	1	−2
−2	1	−2	−1
2	1	0	1

Use a hand calculator or spreadsheet to answer the following questions:

a. Calculate the observations in terms of deviations from their means. That is, find $x_{i2}^* = x_{i2} - \bar{x}_2$, $x_{i3}^* = x_{i3} - \bar{x}_3$, and $y_i^* = y_i - \bar{y}$.
b. Calculate $\sum y_i^* x_{i2}^*$, $\sum x_{i2}^{*2}$, $\sum y_i^* x_{i3}^*$, $\sum x_{i2}^* x_{i3}^*$, and $\sum x_{i3}^{*2}$.
c. Use the expressions in Appendix 5A to find least squares estimates b_1, b_2, and b_3.
d. Find the least squares residuals $\hat{e}_1$, $\hat{e}_2$, …, $\hat{e}_7$.
e. Find the variance estimate $\hat{\sigma}^2$.
f. Find the sample correlation between x_2 and x_3.
g. Find the standard error for b_2.
h. Find *SSE*, *SST*, *SSR*, and R^2.

5.2 Use your answers to Exercise 5.1 to

a. Compute a 95% interval estimate for β_2.
b. Test the hypothesis $H_0: \beta_2 = 1.25$ against the alternative that $H_1: \beta_2 \neq 1.25$.

5.3 Consider the following model that relates the percentage of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + e$$

This model was estimated using 1200 observations from London. An incomplete version of this output is provided in Table 5.6.

TABLE 5.6 Output for Exercise 5.3

Dependent Variable: *WALC*				
Included observations: 1200				
Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	1.4515	2.2019		0.5099
ln(*TOTEXP*)	2.7648		5.7103	0.0000
NK		0.3695	−3.9376	0.0001
AGE	−0.1503	0.0235	−6.4019	0.0000
R-squared		Mean dependent var		6.19434
S.E. of regression		S.D. dependent var		6.39547
Sum squared resid	46221.62			

a. Fill in the following blank spaces that appear in this table.
- **i.** The *t*-statistic for b_1.
- **ii.** The standard error for b_2.
- **iii.** The estimate b_3.
- **iv.** R^2.
- **v.** $\hat{\sigma}$.

b. Interpret each of the estimates b_2, b_3, and b_4.

c. Compute a 95% interval estimate for β_4.What does this interval tell you?

d. Are each of the coefficient estimates significant at a 5% level? Why?

e. Test the hypothesis that the addition of an extra child decreases the mean budget share of alcohol by 2 percentage points against the alternative that the decrease is not equal to 2 percentage points. Use a 5% significance level.

5.4 Consider the following model that relates the percentage of a household's budget spent on alcohol, *WALC*, to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + \beta_5 AGE^2 + e$$

Some output from estimating this model using 1200 observations from London is provided in Table 5.7. The covariance matrix relates to the coefficients b_3, b_4, and b_5.

a. Find a point estimate and a 95% interval estimate for the change in the mean budget percentage share for alcohol when a household has an extra child.

b. Find a point estimate and a 95% interval estimate for the marginal effect of *AGE* on the mean budget percentage share for alcohol when (i) $AGE = 25$, (ii) $AGE = 50$, and (iii) $AGE = 75$.

c. Find a point estimate and a 95% interval estimate for the age at which the mean budget percentage share for alcohol is at a minimum.

d. Summarize what you have discovered from the point and interval estimates in (a), (b), and (c).

e. Let $\mathbf{X}$ represent all the observations on all the explanatory variables. If $(e|\mathbf{X})$ is normally distributed, which of the above interval estimates are valid in finite samples? Which ones rely on a large sample approximation?

f. If $(e|\mathbf{X})$ is not normally distributed, which of the above interval estimates are valid in finite samples? Which ones rely on a large sample approximation?

TABLE 5.7 Output for Exercise 5.4

	Variable	Coefficient
	C	8.149
	ln($TOTEXP$)	2.884
	NK	−1.217
	AGE	−0.5699
	AGE^2	0.005515

	Covariance matrix		
	NK	AGE	AGE^2
NK	0.1462	−0.01774	0.0002347
AGE	−0.01774	0.03204	−0.0004138
AGE^2	0.0002347	−0.0004138	0.000005438

5.5 For each of the following two time-series regression models, and assuming MR1–MR6 hold, find $\text{var}(b_2|\mathbf{x})$ and examine whether the least squares estimator is consistent by checking whether $\lim_{T\to\infty}\text{var}(b_2|\mathbf{x}) = 0$.

a. $y_t = \beta_1 + \beta_2 t + e_t, t = 1, 2, \ldots, T$. Note that $\mathbf{x} = (1, 2, \ldots, T)$, $\sum_{t=1}^{T}(t - \bar{t})^2 = \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2 \Big/ T$, $\sum_{t=1}^{T} t = T(T+1)/2$ and $\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6$.

b. $y_t = \beta_1 + \beta_2(0.5)^t + e_t, t = 1, 2, \ldots, T$. Here, $\mathbf{x} = \left(0.5, 0.5^2, \ldots, 0.5^T\right)$. Note that the sum of a geometric progression with first term r and common ratio r is

$$S = r + r^2 + r^3 + \cdots + r^n = \frac{r(1 - r^n)}{1 - r}$$

c. Provide an intuitive explanation for these results.

5.6 Suppose that, from a sample of 63 observations, the least squares estimates and the corresponding estimated covariance matrix are given by

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix} \qquad \widehat{\text{cov}}(b_1, b_2, b_3) = \begin{bmatrix} 3 & -2 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

Using a 5% significance level, and an alternative hypothesis that the equality does not hold, test each of the following null hypotheses:

a. $\beta_2 = 0$

b. $\beta_1 + 2\beta_2 = 5$

c. $\beta_1 - \beta_2 + \beta_3 = 4$

5.7 After estimating the model $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$ with $N = 203$ observations, we obtain the following information: $\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2 = 1780.7$, $\sum_{i=1}^{N}(x_{i3} - \bar{x}_3)^2 = 3453.3$, $b_2 = 0.7176$, $b_3 = 1.0516$, $SSE = 6800.0$, and $r_{23} = 0.7087$.

a. What are the standard errors of the least squares estimates b_2 and b_3?

b. Using a 5% significance level, test the hypothesis $H_0: \beta_2 = 0$ against the alternative $H_1: \beta_2 \neq 0$.

c. Using a 10% significance level, test the hypothesis $H_0: \beta_3 \leq 0.9$ against the alternative $H_1: \beta_3 > 0.9$.

d. Given that $\widehat{\text{cov}}(b_2, b_3) = -0.019521$, use a 1% significance level to test the hypothesis $H_0: \beta_2 = \beta_3$ against the alternative $H_1: \beta_2 \neq \beta_3$.

5.8 There were 79 countries who competed in the 1996 Olympics and won at least one medal. For each of these countries, let *MEDALS* be the total number of medals won, *POPM* be population in millions,

and *GDPB* be GDP in billions of 1995 dollars. Using these data we estimate the regression model $MEDALS = \beta_1 + \beta_2 POPM + \beta_3 GDPB + e$ to obtain

$$\widehat{MEDALS} = 5.917 + 0.01813\,POPM + 0.01026\,GDPB \qquad R^2 = 0.4879$$
$$\text{(se)} \qquad (1.510) \quad (0.00819) \qquad\quad (0.00136)$$

a. Given assumptions MR1–MR6 hold, interpret the coefficient estimates for β_2 and β_3.
b. Interpret R^2.
c. Using a 1% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?
d. Using a 1% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?
e. Test the following hypotheses using a 5% significance level:
  i. $H_0: \beta_2 = 0.01$ against the alternative $H_1: \beta_2 \neq 0.01$
  ii. $H_0: \beta_2 = 0.02$ against the alternative $H_1: \beta_2 \neq 0.02$
  iii. $H_0: \beta_2 = 0.03$ against the alternative $H_1: \beta_2 \neq 0.03$
  iv. $H_0: \beta_2 = 0.04$ against the alternative $H_1: \beta_2 \neq 0.04$
  Are these test results contradictory? Why or why not?
f. Find a 95% interval estimate for β_2 and comment on it.

5.9 There were 64 countries who competed in the 1992 Olympics and won at least one medal. For each of these countries, let *MEDALS* be the total number of medals won, *POPM* be population in millions, and *GDPB* be GDP in billions of 1995 dollars. Excluding the United Kingdom, and using $N = 63$ observations, the model $MEDALS = \beta_1 + \beta_2 \ln(POPM) + \beta_3 \ln(GDPB) + e$ was estimated as

$$\widehat{MEDALS} = -13.153 + 2.764 \ln(POPM) + 4.270 \ln(GDPB) \qquad R^2 = 0.275$$
$$\text{(se)} \qquad (5.974) \quad (2.070) \qquad\quad (1.718)$$

a. Given assumptions MR1–MR6 hold, interpret the coefficient estimates for β_2 and β_3.
b. Interpret R^2.
c. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?
d. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?
e. Use the model to find point and 95% interval estimates for the expected number of medals won by the United Kingdom whose population and GDP in 1992 were 58 million and $1010 billion, respectively. [The standard error for $b_1 + \ln(58) \times b_2 + \ln(1010) \times b_3$ is 4.22196.]
f. The United Kingdom won 20 medals in 1992. Is the model a good one for predicting the mean number of medals for the United Kingdom? What is an approximate *p*-value for a test of $H_0: \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 = 20$ versus $H_1: \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 \neq 20$?
g. Without doing any of the calculations, write down the expression that is used to compute the standard error given in part (e).

5.10 Using data from 1950 to 1996 ($T = 47$ observations), the following equation for explaining wheat yield in the Mullewa Shire of Western Australia was estimated as

$$\widehat{YIELD}_t = 0.1717 + 0.01117t + 0.05238\,RAIN_t$$
$$\text{(se)} \qquad (0.1537) \quad (0.00262) \quad (0.01367)$$

where $YIELD_t$ = wheat yield in tonnes per hectare in year t;
$TREND_t$ is a trend variable designed to capture technological change, with observations $t = 1, 2, \ldots, 47$;
$RAIN_t$ is total rainfall in inches from May to October (the growing season) in year t. The sample mean and standard deviation for *RAIN* are $\bar{x}_{RAIN} = 10.059$ and $s_{RAIN} = 2.624$.

a. Given assumptions MR1–MR5 hold, interpret the estimates for the coefficients of t and $RAIN$.
b. Using a 5% significance level, test the null hypothesis that technological change increases mean yield by no more than 0.01 tonnes per hectare per year against the alternative that the mean yield increase is greater than 0.01.
c. Using a 5% significance level, test the null hypothesis that an extra inch of rainfall increases mean yield by 0.03 tonnes per hectare against the alternative that the increase is not equal to 0.03.
d. Adding $RAIN^2$ to the equation and reestimating yields

$$\begin{array}{lcccc} \widehat{YIELD}_t = & -0.6759 & +0.011671t & +0.2229RAIN_t & -0.008155RAIN_t^2 \\ \text{(se)} & (0.3875) & (0.00250) & (0.0734) & (0.003453) \end{array}$$

What is the rationale for including $RAIN^2$? Does it have the expected sign?
e. Repeat part (b) using the model estimated in (d).
f. Repeat part (c) using the model estimated in (d), testing the hypothesis at the mean value of rainfall. (The estimated covariance between b_3 and b_4 (the coefficients of $RAIN$ and $RAIN^2$) is $\widehat{\text{cov}}(b_3, b_4) = -0.0002493$.)
g. Use the model in (d) to forecast yield in 1997, when the rainfall was 9.48 inches.
h. Suppose that you wanted to forecast 1997 yield before the rainfall was observed. What would be your forecast from the model in (a)? What would it be from the model in (d)?

5.11 When estimating wage equations, we expect that young, inexperienced workers will have relatively low wages; with additional experience their wages will rise, but then begin to decline after middle age, as the worker nears retirement. This life-cycle pattern of wages can be captured by introducing experience and experience squared to explain the level of wages. If we also include years of education, we have the equation

$$WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

a. What is the marginal effect of experience on the mean wage?
b. What signs do you expect for each of the coefficients β_2, β_3, and β_4? Why?
c. After how many years of experience does the mean wage start to decline? (Express your answer in terms of β's.)
d. Estimating this equation using 600 observations yields

$$\begin{array}{lcccc} \widehat{WAGE} = & -16.308 & +2.329EDUC & +0.5240EXPER & -0.007582EXPER^2 \\ \text{(se)} & (2.745) & (0.163) & (0.1263) & (0.002532) \end{array}$$

The estimated covariance between b_3 and b_4 is $\widehat{\text{cov}}(b_3, b_4) = -0.00030526$. Find 95% interval estimates for the following:
i. The marginal effect of education on mean wage
ii. The marginal effect of experience on mean wage when $EXPER = 4$
iii. The marginal effect of experience on mean wage when $EXPER = 25$
iv. The number of years of experience after which the mean wage declines

5.12 This exercise uses data on 850 houses sold in Baton Rouge, Louisiana during mid-2005. We will be concerned with the selling price in thousands of dollars ($PRICE$), the size of the house in hundreds of square feet ($SQFT$), and the age of the house in years (AGE). The following two regression models were estimated:

$$PRICE = \alpha_1 + \alpha_2 AGE + v \quad \text{and} \quad SQFT = \delta_1 + \delta_2 AGE + u$$

The sums of squares and sums of cross products of the residuals from estimating these two equations are $\sum_{i=1}^{850} \hat{v}_i^2 = 10377817$, $\sum_{i=1}^{850} \hat{u}_i^2 = 75773.4$, $\sum_{i=1}^{850} \hat{u}_i \hat{v}_i = 688318$.
a. Find the least-squares estimate of β_2 in the model $PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + e$.
b. Let $\hat{e}_i = \hat{v}_i - b_2 \hat{u}_i$. Show that $\sum_{i=1}^{850} \hat{e}_i^2 = \sum_{i=1}^{850} \hat{v}_i^2 - b_2 \sum_{i=1}^{850} \hat{v}_i \hat{u}_i$ where b_2 is the least-squares estimate for β_2.
c. Find an estimate of $\sigma^2 = \text{var}(e_i)$.
d. Find the standard error for b_2.
e. What is an approximate p-value for testing $H_0: \beta_2 \geq 9.5$ against the alternative $H_1: \beta_2 < 9.5$? What do you conclude from this p-value?

5.13 A concept used in macroeconomics is Okun's Law, which states that the change in unemployment from one period to the next depends on the rate of growth of the economy relative to a "normal" growth rate:

$$U_t - U_{t-1} = -\gamma(G_t - G_N)$$

where U_t is the unemployment rate in period t, G_t is the growth rate in period t, the "normal" growth rate G_N is that which is required to maintain a constant rate of unemployment, and $0 < \gamma < 1$ is an adjustment coefficient.

a. Show that the model can be written as $DU_t = \beta_1 + \beta_2 G_t$, where $DU_t = U_t - U_{t-1}$ is the change in the unemployment rate, $\beta_1 = \gamma G_N$, and $\beta_2 = -\gamma$.

b. Estimating this model with quarterly seasonally adjusted U.S. data from 1970 Q1 to 2014 Q4 yields

$$\widehat{DU}_t = 0.1989 - 0.2713 G_t \qquad \hat{\sigma} = 0.2749$$

$$\operatorname{cov}(b_1, b_2) = \begin{pmatrix} 0.0007212 & -0.0004277 \\ -0.0004277 & 0.0006113 \end{pmatrix}$$

Use the estimates b_1 and b_2 to find estimates $\hat{\gamma}$ and $\hat{G}_N$.

c. Find standard errors for b_1, b_2, $\hat{\gamma}$, and $\hat{G}_N$. Are all these estimates significantly different from zero at a 5% level?

d. Using a 5% significance level test the null hypothesis that the natural growth rate is 0.8% per quarter against the alternative it is not equal to 0.8%.

e. Find a 95% interval estimate for the adjustment coefficient.

f. Find a 95% interval estimate for $E(U_{2015Q1} | U_{2014Q4} = 5.7991, G_{2015Q1} = 0.062)$.

g. Find a 95% prediction interval for U_{2015Q1} given $U_{2014Q4} = 5.7991$ and $G_{2015Q1} = 0.062$. Explain the difference between this interval and that in (f).

5.14 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where the pairs (y_i, x_i), $i = 1, 2, \ldots, N$, are random independent draws from a population.

a. Suppose the marginal distribution of x_i is log-normal. To appreciate the nature of the log-normal distribution, consider a normal random variable $W \sim N(\mu_W, \sigma_W^2)$. Then, $X = e^W$ has a log-normal distribution with mean $\mu_X = \exp(\mu_W + \sigma_W^2/2)$ and variance $\sigma_X^2 = (\exp(\sigma_W^2) - 1)\mu_X^2$. Assume that $(e_i | x_i) \sim N(0, \sigma_e^2)$.

i. Will the least squares estimator (b_1, b_2) for the parameters (β_1, β_2) be unbiased?

ii. Will it be consistent?

iii. Will it be normally distributed conditional on $\mathbf{x} = (x_1, x_2, \ldots, x_N)$?

iv. Will the marginal distribution of (b_1, b_2) (not conditional on $\mathbf{x}$) be normally distributed?

v. Will t-tests for β_1 and β_2 be justified in finite samples or are they only large sample approximations?

vi. Suppose $\mu_w = 0$, $\sigma_w^2 = 1$, and $x_i = \exp(w_i)$. What is the asymptotic variance of the least squares estimator for β_2? (Express in terms of σ_e^2 and N.)

b. Suppose now that $x_i \sim N(0, 1)$ and that $(e_i | x_i)$ has a log-normal distribution with mean and variance $\mu_e = \exp(\mu_v + \sigma_v^2/2)$ and $\sigma_e^2 = (\exp(\sigma_v^2) - 1)\mu_e^2$, where $v = \ln(e) \sim N(\mu_v, \sigma_v^2)$.

i. Show that we can rewrite the model as $y_i = \beta_1^* + \beta_2 x_i + e_i^*$ where

$$\beta_1^* = \beta_1 + \exp(\mu_v + \sigma_v^2/2) \text{ and } e_i^* = e_i - \exp(\mu_v + \sigma_v^2/2)$$

ii. Show that $E(e_i^* | x_i) = 0$ and $\operatorname{var}(e_i^* | x_i) = \sigma_e^2$.

iii. Will the least squares estimator b_2 for the parameter β_2 be unbiased?

iv. Will it be consistent?

v. Will it be normally distributed conditional on $\mathbf{x} = (x_1, x_2, \ldots, x_N)$?

vi. Will the marginal distribution of b_2 (not conditional on $\mathbf{x}$) be normally distributed?

vii. Will t-tests for β_2 be justified in finite samples or are they only large sample approximations?

viii. What is the asymptotic variance of the least squares estimator for β_2? (Express in terms of σ_e^2 and N.)

5.15 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where the pairs (y_i, x_i), $i = 1, 2, \ldots, N$, are random independent draws from a population, $x_i \sim N(0, 1)$, and $E(e_i | x_i) = c(x_i^2 - 1)$ where c is a constant.

a. Show that $E(e_i) = 0$.

b. Using the result $\operatorname{cov}(e_i, x_i) = E_x[(x_i - \mu_x) E(e_i | x_i)]$, show that $\operatorname{cov}(e_i, x_i) = 0$.

c. Will the least squares estimator for β_2 be (i) unbiased? (ii) consistent?

5.16 Consider a log-linear regression for the weekly sales of a national brand of canned tuna (brand *A*), expressed as thousands of cans, *CANS*, as a function of the prices of two competing brands (brands *B* and *C*), expressed as percentages of the price of brand *A*. That is,

$$\ln(CANS) = \beta_1 + \beta_2 RPRCE\_B + \beta_3 RPRCE\_C + e$$

where $RPRCE\_B = \left(PRICE_B/PRICE_A\right) \times 100$ and $RPRCE\_C = \left(PRICE_C/PRICE_A\right) \times 100$.

a. Given assumptions MR1–MR5 hold, how do you interpret β_2 and β_3? What signs do you expect for these coefficients? Why?

Using $N = 52$ weekly observations, the least squares estimated equation is

$$\widehat{\ln(CANS)} = -2.724 + 0.0146 RPRCE\_B + 0.02649 RPRCE\_C \qquad \hat{\sigma} = 0.5663$$
$$\text{(se)} \qquad (0.582) \quad (0.00548) \qquad\qquad (0.00544) \qquad\qquad \widehat{\text{cov}}\left(b_2, b_3\right) = -0.0000143$$

b. Using a 10% significance level, test the null hypothesis that an increase in *RPRCE_B* of one percentage point leads to a 2.5% increase in the mean number of cans sold against the alternative that the increase is not 2.5%.

c. Using a 10% significance level, test the null hypothesis that an increase in *RPRCE_C* of one percentage point leads to a 2.5% increase in the mean number of cans sold against the alternative that the increase is not 2.5%.

d. Using a 10% significance level, test $H_0: \beta_2 = \beta_3$ against the alternative $H_1: \beta_2 \neq \beta_3$. Does the outcome of this test contradict your findings from parts (b) and (c)?

e. Which brand do you think is the closer substitute for brand *A*, brand *B*, or brand *C*? Why?

f. Use the corrected predictor introduced in Section 4.5.3 to estimate the expected number of brand *A* cans sold under the following scenarios:

i. $RPRCE\_B = 125$, $RPRCE\_C = 100$

ii. $RPRCE\_B = 111.11$, $RPRCE\_C = 88.89$

iii. $RPRCE\_B = 100$, $RPRCE\_C = 80$

g. The producers of brands *B* and *C* have set the prices of their cans of tuna to be \$1 and 80 cents, respectively. The producer of brand *A* is considering three possible prices for her cans: 80 cents, 90 cents, or \$1. Use the results from part (f) to find which of these three price settings will maximize revenue from sales.

5.8.2 Computer Exercises

5.17 Use econometric software to verify your answers to Exercise 5.1, parts (c), (e), (f), (g), and (h).

5.18 Consider the following two expenditure share equations where the budget share for food *WFOOD*, and the budget share for clothing *WCLOTH*, are expressed as functions of total expenditure *TOTEXP*.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e_F \tag{XR5.18.1}$$

$$WCLOTH = \alpha_1 + \alpha_2 \ln(TOTEXP) + e_C \tag{XR5.18.2}$$

a. A commodity is regarded as a luxury if the coefficient of $\ln(TOTEXP)$ is positive and a necessity if it is negative. What signs would you expect for β_2 and α_2?

b. Using the data in the file *london5*, estimate the above equations using observations on households with one child. Comment on the estimates and their significance. Can you explain any possibly counterintuitive outcomes?

c. Using a 1% significance level, test $H_0: \beta_2 \geq 0$ against the alternative $H_1: \beta_2 < 0$. Why might you set up the hypotheses in this way?

d. Using a 1% significance level, test $H_0: \alpha_2 \geq 0$ against the alternative $H_1: \alpha_2 < 0$. Why might you set up the hypotheses in this way?

e. Estimate the two equations using observations on households with two children. Construct 95% interval estimates for β_2 and α_2 for both one- and two-child households. Based on these interval estimates, would you conjecture that the coefficients of $\ln(TOTEXP)$ are the same or different for one- and two-child households.

f. Use all observations to estimate the following two equations and test, at a 95% significance level, whether your conjectures in part (e) are correct. (NK = number of children in the household.)

$$WFOOD = \gamma_1 + \gamma_2 \ln(TOTEXP) + \gamma_3 NK + \gamma_4 NK \times \ln(TOTEXP) + e_F \quad \text{(XR5.18.3)}$$

$$WCLOTH = \delta_1 + \delta_2 \ln(TOTEXP) + \delta_3 NK + \delta_4 NK \times \ln(TOTEXP) + e_C \quad \text{(XR5.18.4)}$$

g. Compare the estimates for $\partial E(WFOOD|\mathbf{X})/\partial \ln(TOTEXP)$ from (XR5.18.1) for $NK = 1, 2$ with those from (XR5.18.3) for $NK = 1, 2$.

5.19 Consider the following expenditure share equation where the budget share for food $WFOOD$ is expressed as a function of total expenditure $TOTEXP$.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e_F \quad \text{(XR5.19.1)}$$

In Exercise 4.12, it was noted that the elasticity of expenditure on food with respect to total expenditure is given by

$$\varepsilon = 1 + \frac{\beta_2}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

Also, in Exercise 5.18 it was indicated that a good is a necessity if $\beta_2 < 0$.

a. Show that $\beta_2 < 0$ if and only if $\varepsilon < 1$. That is, a good is a necessity if its expenditure elasticity is less than one (inelastic).

b. Use observations in the data file *london5* to estimate (XR5.19.1) and comment on the results.

c. Find point estimates and 95% interval estimates for the mean budget share for food, for total expenditure values (i) $TOTEXP = 50$ (the fifth percentile of $TOTEXP$), (ii) $TOTEXP = 90$ (the median), and (iii) $TOTEXP = 170$ (the 95th percentile).

d. Find point estimates and 95% interval estimates for the elasticity ε, for total expenditure values (i) $TOTEXP = 50$ (the fifth percentile), (ii) $TOTEXP = 90$ (the median), and (iii) $TOTEXP = 170$ (the 95th percentile).

e. Comment on how the mean budget share and the expenditure elasticity for food change as total expenditure changes. How does the reliability of estimation change as total expenditure changes?

5.20 A generalized version of the estimator for β_2 proposed in Exercise 2.9 by Professor I.M. Mean for the regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, 2, \ldots, N$ is

$$\hat{\beta}_{2,mean} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}$$

where $(\bar{y}_1, \bar{x}_1)$ and $(\bar{y}_2, \bar{x}_2)$ are the sample means for the first and second halves of the sample observations, respectively, after ordering the observations according to increasing values of x. Given that assumptions MR1–MR6 hold:

a. Show that $\hat{\beta}_{2,mean}$ is unbiased.

b. Derive an expression for $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$.

c. Write down an expression for $\text{var}\left(\hat{\beta}_{2,mean}\right)$.

d. Under what conditions will $\hat{\beta}_{2,mean}$ be a consistent estimator for β_2?

e. Randomly generate observations on x from a uniform distribution on the interval (0,10) for sample sizes $N = 100, 500, 1000$, and, if your software permits, $N = 5000$. Assuming $\sigma^2 = 1000$, for each sample size, compute:

i. $\text{var}(b_2|\mathbf{x})$ and $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$ where b_2 is the OLS estimator.

ii. Estimates for $E\left[\left(s_x^2\right)^{-1}\right]$ and $E\left[4/\left(\bar{x}_2 - \bar{x}_1\right)^2\right]$ where s_x^2 is the sample standard deviation for x using N as a divisor.

f. Comment on the relative magnitudes of your answers in part (e), (i) and (ii) and how they change as sample size increases. Does it appear that $\hat{\beta}_{2,mean}$ is consistent?

g. Show that $E\left[\left(s_x^2\right)^{-1}\right] \xrightarrow{p} 0.12$ and $E\left[4/\left(\bar{x}_2 - \bar{x}_1\right)^2\right] \xrightarrow{p} 0.16$. [*Hint*: The variance of a uniform random variable defined on the interval (a, b) is $(b-a)^2/12$.]

h. Suppose that the observations on x were not ordered according to increasing magnitude but were randomly assigned to any position. Would the estimator $\hat{\beta}_{2,mean}$ be consistent? Why or why not?

5.21 Using the data in the file *toody5*, estimate the model

$$Y_t = \beta_1 + \beta_2 TREND_t + \beta_3 RAIN_t + \beta_4 RAIN_t^2 + \beta_5 (RAIN_t \times TREND_t) + e_t$$

where Y_t = wheat yield in tons per hectare in the Toodyay Shire of Western Australia in year t; $TREND_t$ is a trend variable designed to capture technological change, with observations 0, 0.1, 0.2, …, 4.7; 0 is for the year 1950, 0.1 is for the year 1951, and so on up to 4.7 for the year 1997; $RAIN_t$ is total rainfall in decimeters (dm) from May to October (the growing season) in year t (1 decimeter = 4 inches).

a. Report your estimates, standard errors, t-values, and p-values in a table.
b. Are each of your estimates significantly different from zero at a (i) 5% level, (ii) 10% level?
c. Do the coefficients have the expected signs? Why? (One of the objectives of technological improvements is the development of drought-resistant varieties of wheat.)
d. Find point and 95% interval estimates of the marginal effect of extra rainfall in (i) 1959 when the rainfall was 2.98 dm and (ii) 1995 when the rainfall was 4.797 dm. Discuss the results.
e. Find point and 95% interval estimates for the amount of rainfall that would maximize expected yield in (i) 1959 and (ii) 1995. Discuss the results.

5.22 Using the data in the file *toody5*, estimate the model

$$Y_t = \beta_1 + \beta_2 TREND_t + \beta_3 RAIN_t + \beta_4 RAIN_t^2 + \beta_5 (RAIN_t \times TREND_t) + e_t$$

where Y_t = wheat yield in tons per hectare in the Toodyay Shire of Western Australia in year t; $TREND_t$ is a trend variable designed to capture technological change, with observations 0, 0.1, 0.2, …, 4.7; 0 is for the year 1950, 0.1 is for the year 1951, and so on up to 4.7 for the year 1997; $RAIN_t$ is total rainfall in decimeters (dm) from May to October (the growing season) in year t (1 decimeter = 4 inches).

a. Report your estimates, standard errors, t-values, and p-values in a table.
b. For 1974, when $TREND = 2.4$ and $RAIN = 4.576$, use a 5% significance level to test the null hypothesis that extra rainfall will not increase expected yield against the alternative that it will increase expected yield.
c. Assuming rainfall is equal to its median value of 3.8355 dm, find point and 95% interval estimates of the expected improvement in wheat yield from technological change over the period 1960–1995.
d. There is concern that climate change is leading to a decline in rainfall over time. To test this hypothesis, estimate the equation $RAIN = \alpha_1 + \alpha_2 TREND + e$. Test, at a 5% significance level, the null hypothesis that mean rainfall is not declining over time against the alternative hypothesis that it is declining.
e. Using the estimated equation from part (d), estimate mean rainfall in 1960 and in 1995.
f. Suppose that $TREND_{1995} = TREND_{1960}$, implying there had been no technological change from 1960 to 1995. Use the estimates from part (e) to find an estimate of the decline in mean yield from 1960 to 1995 attributable to climate change.
g. Suppose that $E(RAIN_{1995}) = E(RAIN_{1960})$, implying there had been no rainfall change from 1960 to 1995. Find an estimate of the increase in mean yield from 1960 to 1995 attributable to technological change.
h. Compare the estimates you obtained in parts (c), (f), and (g).

5.23 The file *cocaine* contains 56 observations on variables related to sales of cocaine powder in northeastern California over the period 1984–1991. The data are a subset of those used in the study Caulkins, J. P. and R. Padman (1993), "Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748–757. The variables are

PRICE = price per gram in dollars for a cocaine sale
QUANT = number of grams of cocaine in a given sale
QUAL = quality of the cocaine expressed as percentage purity
TREND = a time variable with 1984 = 1 up to 1991 = 8
Consider the regression model

$$PRICE = \beta_1 + \beta_2 QUANT + \beta_3 QUAL + \beta_4 TREND + e$$

a. What signs would you expect on the coefficients β_2, β_3, and β_4?

b. Use your computer software to estimate the equation. Report the results and interpret the coefficient estimates. Have the signs turned out as you expected?

c. What proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time?

d. It is claimed that the greater the number of sales, the higher the risk of getting caught. Thus, sellers are willing to accept a lower price if they can make sales in larger quantities. Set up H_0 and H_1 that would be appropriate to test this hypothesis. Carry out the hypothesis test.

e. Test the hypothesis that the quality of cocaine has no influence on expected price against the alternative that a premium is paid for better-quality cocaine.

f. What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction?

5.24 The file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana during 2009–2013. We will be concerned with the selling price in thousands of dollars (*PRICE*), the size of the house in hundreds of square feet (*SQFT*), and the age of the house measured as a categorical variable (*AGE*), with 1 representing the newest and 11 the oldest. Let **X** denote all observations on *SQFT* and *AGE*. Use all observations to estimate the following regression model:

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 (SQFT \times AGE) + e$$

a. Report the results. Are the estimated coefficients significantly different from zero?

b. Write down expressions for the marginal effects $\partial E(PRICE|\mathbf{X})/\partial SQFT$ and $\partial E(PRICE|\mathbf{X})/\partial AGE$. Interpret each of the coefficients. Given the categorical nature of the variable *AGE*, what assumptions are being made?

c. Find point and 95% interval estimates for the marginal effect $\partial E(PRICE|\mathbf{X})/\partial SQFT$ for houses that are (i) 5 years old, (ii) 20 years old, and (iii) 40 years old. How do these estimates change as *AGE* increases? (Refer to the file *collegetown.def* for the definition of *AGE*.)

d. As a house gets older and moves from one age category to the next, the expected price declines by \$6000. Set up this statement as a null hypothesis for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet. Using a 5% significance level, test each of the null hypotheses against an alternative that the price decline is not \$6000. Discuss the outcomes.

e. Find a 95% prediction interval for the price of a 60-year old house with 2500 square feet. In the data set there are four 60-year old houses with floor space between 2450 and 2550 square feet. What prices did they sell for? How many of these prices fall within your prediction interval? Is the model a good one for forecasting price?

5.25 The file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana during 2009–2013. We will be concerned with the selling price in thousands of dollars (*PRICE*), and the size of the house in hundreds of square feet (*SQFT*). Use all observations to estimate the following regression model:

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2} + e$$

Suppose that assumptions MR1–MR6 all hold. In particular, $(e|SQFT) \sim N\left(0, \sigma^2\right)$.

a. Report the results. Are the estimated coefficients significantly different from zero?

b. Write down an expression for the marginal effect $\partial E\left[\ln(PRICE|SQFT)\right]/\partial SQFT$. Discuss the nature of this marginal effect and the expected signs for β_2 and β_3.

c. Find and interpret point and 95% interval estimates for $\partial E\left[\ln(PRICE|SQFT)\right]/\partial SQFT$ for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet.

d. Show that

$$\frac{\partial E[PRICE|SQFT]}{\partial SQFT} = \left(\beta_2 + \frac{\beta_3}{2SQFT^{1/2}}\right) \times \exp\left\{\beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2} + \sigma^2/2\right\}$$

For future reference, we write this expression as $\partial E(PRICE|SQFT)/\partial SQFT = S \times C$ where

$$S = \left(\beta_2 + \frac{\beta_3}{2SQFT^{1/2}}\right) \times \exp\left\{\beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2}\right\} \quad \text{and } C = \exp\left\{\sigma^2/2\right\}$$

Correspondingly, we let $\hat{S}$ and $\hat{C}$ denote estimates for S and C obtained by replacing unknown parameters by their estimates.

e. Estimate $\partial E(PRICE|SQFT)/\partial SQFT = S \times C$ for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet.

f. Finding the asymptotic standard errors for the estimates in (e) is tricky because of the product $\hat{S} \times \hat{C}$. To avoid such trickiness, find the standard errors for $\hat{S}$ for each type of house in (e).

g. For each type of house, and a 5% significance level, use the estimates from (e) and the standard errors from (f) to test the hypotheses

$$H_0: \frac{\partial E(PRICE|SQFT)}{\partial SQFT} = 9 \quad H_1: \frac{\partial E(PRICE|SQFT)}{\partial SQFT} \neq 9$$

What do you conclude?

h. (optional) To get the "correct" standard errors for $\hat{S} \times \hat{C}$, we proceed as follows. First, given $\widehat{\text{var}}\left(\hat{\sigma}^2\right) = 2\sigma^4/(N-K)$, find an estimate for $\widehat{\text{var}}(\hat{C})$. It can be shown that $\hat{S}$ and $\hat{C}$ are independent. Using results on the product of independent random variables, an estimator for the variance of $\hat{S} \times \hat{C}$ is

$$\widehat{\text{var}}(\hat{S} \times \hat{C}) = \widehat{\text{var}}\left(\widehat{\frac{\partial E(PRICE|SQFT)}{\partial SQFT}}\right) = \hat{S}^2\widehat{\text{var}}(\hat{C}) + \hat{C}^2\widehat{\text{var}}(\hat{S}) + \widehat{\text{var}}(\hat{C})\,\widehat{\text{var}}(\hat{S})$$

Use this result to find standard errors for $\hat{S} \times \hat{C}$. How do they compare with the standard errors obtained in (f)? Are they likely to change the outcomes of the hypothesis tests in (g)?

5.26 Consider the presidential voting data (data file *fair5*) introduced in Exercise 2.23. Details of the variables can be found in that exercise.

a. Using all observations, estimate the regression model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLAT + e$$

Report the results. Are the estimates for β_2 and β_3 significantly different from zero at a 10% significance level? Did you use one- or two-tail tests? Why?

b. Assume the inflation rate is 3% and the Democrats are the incumbent party. Predict the percentage vote for both parties when the growth rate is (i) −2%, (ii) 0%, and (iii) 3%.

c. Assume the inflation rate is 3% and the Republicans are the incumbent party. Predict the percentage vote for both parties when the growth rate is (i) −2%, (ii) 0%, and (iii) 3%.

d. Based on your answers to parts (b) and (c), do you think the popular vote tends to be more biased toward the Democrats or the Republicans?

e. Consider the following two scenarios:

1. The Democrats are the incumbent party, the growth rate is 2% and the inflation rate is 2%.
2. The Republicans are the incumbent party, the growth rate is 2% and the inflation rate is 2%.

Using a 5% significance level, test the null hypothesis that the expected share of the Democratic vote under scenario 1 is equal to the expected share of the Republican vote under scenario 2.

5.27 In this exercise, we consider the auction market for art first introduced in Exercise 2.24. The variables in the data file *ashcan_small* that we will be concerned with are as follows:

RHAMMER = the price at which a painting sold in thousands of dollars
YEARS_OLD = the time between completion of the painting and when it was sold
INCHSQ = the size of the painting in square inches

Create a new variable $INCHSQ10 = INCHSQ/10$ to express size in terms of tens of square inches. Only consider observations where the art was sold ($SOLD = 1$).

a. Estimate the following equation and report the results:

$$RHAMMER = \beta_1 + \beta_2 YEARS\_OLD + \beta_3 INCHSQ10 + e$$

b. How much do paintings appreciate on a yearly basis? Find a 95% interval estimate for the expected yearly price increase.

c. How much more valuable are large paintings? Find a 95% interval estimate for the expected extra value from an extra 10 square inches.

d. Add the variable $INCHSQ10^2$ to the model and re-estimate. Report the results. Why would you consider adding this variable?

e. Does adding this variable have much impact on the interval estimate in part (b)?

f. Find 95% interval estimates for the expected extra value from an extra 10 square inches for art of the following sizes: (i) 50 square inches (sixth percentile), (ii) 250 square inches (approximately the median), and (iii) 900 square inches (97th percentile). Comment on how the value of an extra 10 square inches changes as the painting becomes larger.

g. Find a 95% interval estimate for the painting size that maximizes price.

h. Find a 95% interval estimate for the expected price of a 75-year-old, 100-square-inch painting.

i. How long would you have to keep a 100-square-inch painting for the expected price to become positive?

5.28 In this exercise, we consider the auction market for art first introduced in Exercise 2.24. The variables in the data file *ashcan_small* that we will be concerned with are as follows:

RHAMMER = the price at which a painting sold in thousands of dollars
YEARS_OLD = the time between completion of the painting and when it was sold
INCHSQ = the size of the painting in square inches

Create a new variable $INCHSQ10 = INCHSQ/10$ to express size in terms of tens of square inches. Only consider observations where the art was sold ($SOLD = 1$).

a. Estimate the following log-linear equation and report the results:

$$\ln(RHAMMER) = \beta_1 + \beta_2 YEARS\_OLD + \beta_3 INCHSQ10 + e$$

b. How much do paintings appreciate on a yearly basis? Find a 95% interval estimate for the expected percentage price increase per year.

c. How much more valuable are large paintings? Using a 5% significance level, test the null hypothesis that painting an extra 10 square inches increases the value by 2% or less against the alternative that it increases the value by more than 2%.

d. Add the variable $INCHSQ10^2$ to the model and re-estimate. Report the results. Why would you consider adding this variable?

e. Does adding this variable have much impact on the interval estimate in part (b)?

f. Redo the hypothesis test in part (c) for art of the following sizes: (i) 50 square inches (sixth percentile), (ii) 250 square inches (approximately the median), and (iii) 900 square inches (97th percentile). What do you observe?

g. Find a 95% interval estimate for the painting size that maximizes price.

h. Find a 95% interval estimate for the expected price of a 75-year-old, 100-square-inch painting. (Use the estimator $\exp\left\{E\left[\ln(RHAMMER|YEARS\_OLD = 75, INCHSQ10 = 10)\right]\right\}$ and its standard error.)

5.29 What is the relationship between crime and punishment? This important question has been examined by Cornwell and Trumbull[16] using a panel of data from North Carolina. The cross sections are 90 counties, and the data are annual for the years 1981–1987. The data are in the file *crime*.

Using the data from 1986, estimate a regression relating the log of the crime rate *LCRMRTE* to the probability of an arrest *PRBARR* (the ratio of arrests to offenses), the probability of conviction *PRBCONV* (the ratio of convictions to arrests), the probability of a prison sentence *PRBPRIS* (the ratio of prison sentences to convictions), the number of police per capita *POLPC*, and the weekly wage in construction *WCON*. Write a report of your findings. In your report, explain what effect you would expect each of the variables to have on the crime rate and note whether the estimated coefficients have the expected signs and are significantly different from zero. What variables appear to be the most important for crime deterrence? Can you explain the sign for the coefficient of *POLPC*?

5.30 In Section 5.7.4, we discovered that the optimal level of advertising for Big Andy's Burger Barn, $ADVERT_0$, satisfies the equation $\beta_3 + 2\beta_4 ADVERT_0 = 1$. Using a 5% significance level, test whether each of the following levels of advertising could be optimal: (a) $ADVERT_0 = 1.75$, (b) $ADVERT_0 = 1.9$, and (c) $ADVERT_0 = 2.3$. What are the *p*-values for each of the tests?

5.31 Each morning between 6:30 AM and 8:00 AM Bill leaves the Melbourne suburb of Carnegie to drive to work at the University of Melbourne. The time it takes Bill to drive to work (*TIME*), depends on the departure time (*DEPART*), the number of red lights that he encounters (*REDS*), and the number of trains that he has to wait for at the Murrumbeena level crossing (*TRAINS*). Observations on these

[16] "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics*, 76, 1994, 360–366. The data were kindly provided by the authors.

variables for the 249 working days in 2015 appear in the file *commute5*. *TIME* is measured in minutes. *DEPART* is the number of minutes after 6:30 AM that Bill departs.

a. Estimate the equation

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

Report the results and interpret each of the coefficient estimates, including the intercept β_1.

b. Find 95% interval estimates for each of the coefficients. Have you obtained precise estimates of each of the coefficients?

c. Using a 5% significance level, test the null hypothesis that Bill's expected delay from each red light is 2 minutes or more against the alternative that it is less than 2 minutes.

d. Using a 10% significance level, test the null hypothesis that the expected delay from each train is 3 minutes against the alternative that it is not 3 minutes.

e. Using a 5% significance level, test the null hypothesis that Bill can expect a trip to be at least 10 minutes longer if he leaves at 7:30 AM instead of 7:00 AM, against the alternative that it will not be 10 minutes longer. (Assume other things are equal.)

f. Using a 5% significance level, test the null hypothesis that the expected delay from a train is at least three times greater than the expected delay from a red light against the alternative that it is less than three times greater.

g. Suppose that Bill encounters six red lights and one train. Using a 5% significance level, test the null hypothesis that leaving Carnegie at 7:00 AM is early enough to get him to the university on or before 7:45 AM against the alternative that it is not. [Carry out the test in terms of the expected time $E(TIME|\mathbf{X})$ where $\mathbf{X}$ represents the observations on all explanatory variables.]

h. Suppose that, in part (g), it is imperative that Bill is not late for his 7:45 AM meeting. Have the null and alternative hypotheses been set up correctly? What happens if these hypotheses are reversed?

5.32 Reconsider the variables and model from Exercise 5.31

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

Suppose that Bill is mainly interested in the magnitude of the coefficient β_2. He has control over his departure time, but no control over the red lights or the trains.

a. Regress *DEPART* on the variables *REDS* and *TRAINS* and save the residuals. Which coefficient estimates are significantly different from zero at a 5% level? For the significant coefficient(s), do you think the relationship is causal?

b. Regress *TIME* on the variables *REDS* and *TRAINS* and save the residuals. Are the estimates for the coefficients of *REDS* and *TRAINS* very different from the estimates for β_3 and β_4 obtained by estimating the complete model with *DEPART* included?

c. Use the residuals from parts (a) and (b) to estimate the coefficient β_2 and adjust the output to obtain its correct standard error.

5.33 Use the observations in the data file *cps5_small* to estimate the following model:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6 (EDUC \times EXPER) + e$$

a. At what levels of significance are each of the coefficient estimates "significantly different from zero"?

b. Obtain an expression for the marginal effect $\partial E\left[\ln(WAGE)|EDUC, EXPER\right]/\partial EDUC$. Comment on how the estimate of this marginal effect changes as *EDUC* and *EXPER* increase.

c. Evaluate the marginal effect in part (b) for all observations in the sample and construct a histogram of these effects. What have you discovered? Find the median, 5th percentile, and 95th percentile of the marginal effects.

d. Obtain an expression for the marginal effect $\partial E\left[\ln(WAGE)|EDUC, EXPER\right]/\partial EXPER$. Comment on how the estimate of this marginal effect changes as *EDUC* and *EXPER* increase.

e. Evaluate the marginal effect in part (d) for all observations in the sample and construct a histogram of these effects. What have you discovered? Find the median, 5th percentile, and 95th percentile of the marginal effects.

f. David has 17 years of education and 8 years of experience, while Svetlana has 16 years of education and 18 years of experience. Using a 5% significance level, test the null hypothesis that Svetlana's expected log-wage is equal to or greater than David's expected log-wage, against the alternative that David's expected log-wage is greater. State the null and alternative hypotheses in terms of the model parameters.

g. After eight years have passed, when David and Svetlana have had eight more years of experience, but no more education, will the test result in (f) be the same? Explain this outcome?

h. Wendy has 12 years of education and 17 years of experience, while Jill has 16 years of education and 11 years of experience. Using a 5% significance level, test the null hypothesis that their marginal effects of extra experience are equal against the alternative that they are not. State the null and alternative hypotheses in terms of the model parameters.

i. How much longer will it be before the marginal effect of experience for Jill becomes negative? Find a 95% interval estimate for this quantity.

Appendix 5A Derivation of Least Squares Estimators

In Appendix 2A, we derived expressions for the least squares estimators b_1 and b_2 in the simple regression model. In this appendix, we proceed with a similar exercise for the multiple regression model; we describe how to obtain expressions for b_1, b_2, and b_3 in a model with two explanatory variables. Given sample observations on y, x_2, and x_3, the problem is to find values for β_1, β_2, and β_3 that minimize

$$S(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3})^2$$

The first step is to partially differentiate S with respect to β_1, β_2, and β_3 and to set the first-order partial derivatives to zero. This yields

$$\frac{\partial S}{\partial \beta_1} = 2N\beta_1 + 2\beta_2 \sum x_{i2} + 2\beta_3 \sum x_{i3} - 2\sum y_i$$

$$\frac{\partial S}{\partial \beta_2} = 2\beta_1 \sum x_{i2} + 2\beta_2 \sum x_{i2}^2 + 2\beta_3 \sum x_{i2}x_{i3} - 2\sum x_{i2}y_i$$

$$\frac{\partial S}{\partial \beta_3} = 2\beta_1 \sum x_{i3} + 2\beta_2 \sum x_{i2}x_{i3} + 2\beta_3 \sum x_{i3}^2 - 2\sum x_{i3}y_i$$

Setting these partial derivatives equal to zero, dividing by 2, and rearranging yields

$$\begin{aligned} Nb_1 + \sum x_{i2}b_2 + \sum x_{i3}b_3 &= \sum y_i \\ \sum x_{i2}b_1 + \sum x_{i2}^2 b_2 + \sum x_{i2}x_{i3}b_3 &= \sum x_{i2}y_i \\ \sum x_{i3}b_1 + \sum x_{i2}x_{i3}b_2 + \sum x_{i3}^2 b_3 &= \sum x_{i3}y_i \end{aligned} \tag{5A.1}$$

The least squares estimators for b_1, b_2, and b_3 are given by the solution of this set of three *simultaneous equations*, known as the **normal equations**. To write expressions for this solution, it is convenient to express the variables as deviations from their means. That is, let

$$y_i^* = y_i - \bar{y}, \quad x_{i2}^* = x_{i2} - \bar{x}_2, \quad x_{i3}^* = x_{i3} - \bar{x}_3$$

Then the least squares estimates b_1, b_2, and b_3 are

$$b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$b_2 = \frac{(\sum y_i^* x_{i2}^*)(\sum x_{i3}^{*2}) - (\sum y_i^* x_{i3}^*)(\sum x_{i2}^* x_{i3}^*)}{(\sum x_{i2}^{*2})(\sum x_{i3}^{*2}) - (\sum x_{i2}^* x_{i3}^*)^2}$$

$$b_3 = \frac{(\sum y_i^* x_{i3}^*)(\sum x_{i2}^{*2}) - (\sum y_i^* x_{i2}^*)(\sum x_{i3}^* x_{i2}^*)}{(\sum x_{i2}^{*2})(\sum x_{i3}^{*2}) - (\sum x_{i2}^* x_{i3}^*)^2}$$

For models with more than three parameters, the solutions become quite messy without using matrix algebra; we will not show them. Computer software used for multiple regression computations solves normal equations such as those in (5A.1) to obtain the least squares estimates.

Appendix 5B The Delta Method

In Sections 3.6, 5.3, 5.4, and 5.5, we discussed estimating and testing **linear combinations** of parameters. If the regression errors are normal, the results discussed there hold in finite samples. If the regression errors are not normal, then those results hold in large samples, as discussed in Section 5.7. We now turn to **nonlinear functions** of regression parameters that were considered in Section 5.7.4 and provide some background for the results given there. You will be surprised in the subsequent chapters how many times we become interested in **nonlinear functions** of regression parameters. For example, we may find ourselves interested in functions such as $g_1(\beta_2) = \exp(\beta_2/10)$ or $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. The first function $g_1(\beta_2)$ is a function of the single parameter β_2. Intuitively, we would estimate this function of β_2 using $g_1(b_2)$. The second function $g_2(\beta_1, \beta_2)$ is a function of two parameters and similarly $g_2(b_1, b_2)$ seems like a reasonable estimator. Working with nonlinear functions of the estimated parameters requires additional tools because, even if the regression errors are normal, nonlinear functions of them are not normally distributed in finite samples, and usual variance formulas do not apply.

5B.1 Nonlinear Function of a Single Parameter

The key to working with nonlinear functions of a single parameter is the Taylor series approximation discussed in Appendix A, Derivative Rule 9. It is stated there as

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x-a) = f(a) + f'(a)(x-a)$$

The value of a function at x is approximately equal to the value of the function at $x = a$, plus the derivative of the function evaluated at $x = a$, times the difference $x - a$. This approximation works well when the function is smooth and the difference $x - a$ is not too large. We will apply this rule to $g_1(b_2)$ replacing x with b_2 and a with β_2

$$g_1(b_2) \cong g_1(\beta_2) + g_1'(\beta_2)(b_2 - \beta_2) \tag{5B.1}$$

This Taylor series expansion of $g_1(b_2)$ shows the following:

1. If $E(b_2) = \beta_2$, then $E[g_1(b_2)] \cong g_1(\beta_2)$.
2. If b_2 is a biased but consistent estimator, so that $b_2 \xrightarrow{p} \beta_2$, then $g_1(b_2) \xrightarrow{p} g_1(\beta_2)$.
3. The variance of $g_1(b_2)$ is given by $\text{var}[g_1(b_2)] \cong [g_1'(\beta_2)]^2 \text{var}(b_2)$, which is known as the **delta method**. The delta method follows from working with the Taylor series approximation

$$\begin{aligned}
\text{var}[g_1(b_2)] &= \text{var}\left[g_1(\beta_2) + g_1'(\beta_2)(b_2 - \beta_2)\right] \\
&= \text{var}\left[g_1'(\beta_2)(b_2 - \beta_2)\right] \text{ because } g_1(\beta_2) \text{ is not random} \\
&= [g_1'(\beta_2)]^2 \text{var}(b_2 - \beta_2) \text{ because } g_1'(\beta_2) \text{ is not random} \\
&= [g_1'(\beta_2)]^2 \text{var}(b_2) \text{ because } \beta_2 \text{ is not random}
\end{aligned}$$

4. The estimator $g_1(b_2)$ has an approximate normal distribution in large samples,

$$g_1(b_2) \overset{a}{\sim} N\left[g_1(\beta_2), [g_1'(\beta_2)]^2 \text{var}(b_2)\right] \tag{5B.2}$$

The asymptotic normality of $g_1(b_2)$ means that we can test nonlinear hypotheses about β_2, such as $H_0: g_1(\beta_2) = c$, and we can construct interval estimates of $g_1(\beta_2)$ in the usual way. To implement the delta method, we replace β_2 by its estimate b_2 and the true variance $\text{var}(b_2)$ by its estimate $\widehat{\text{var}}(b_2)$ which, for the simple regression model, is given in equation (2.21).

EXAMPLE 5.19 | An Interval Estimate for $\exp(\beta_2/10)$

To illustrate the delta method calculations, we use one sample from the $N = 20$ simulation considered in Appendix 5C; it is stored as *mc20*. For these data values, the fitted regression is

$$\hat{y} = 87.44311 + 10.68456x$$
$$\text{(se)}\ (33.8764) \quad (2.1425)$$

The nonlinear function we consider is $g_1(\beta_2) = \exp(\beta_2/10)$. In the simulation we know the value $\beta_2 = 10$ and therefore the value of the function is $g_1(\beta_2) = \exp(\beta_2/10) = e^1 = 2.71828$. To apply the delta method, we need the derivative $g_1'(\beta_2) = \exp(\beta_2/10) \times (1/10)$ (see Appendix A, Derivative Rule 7), and the estimated covariance matrix in Table 5B.1.

The estimated value of the nonlinear function is

$$g_1(b_2) = \exp(b_2/10) = \exp(10.68456/10) = 2.91088$$

The estimated variance is

$$\widehat{\text{var}}[g_1(b_2)] = [g_1'(b_2)]^2 \widehat{\text{var}}(b_2) = [\exp(b_2/10) \times (1/10)]^2 \widehat{\text{var}}(b_2)$$
$$= [\exp(10.68456/10) \times (1/10)]^2 4.59045 = 0.38896$$

TABLE 5B.1 Estimated Covariance Matrix

	b_1	b_2
b_1	1147.61330	−68.85680
b_2	−68.85680	4.59045

and

$$\text{se}[g_1(b_2)] = 0.62367.$$

The 95% interval estimate is

$$g_1(b_2) \pm t_{(0.975,20-2)}\text{se}[g_1(b_2)] = 2.91088 \pm 2.10092 \times 0.62367$$
$$= (1.60061, 4.22116)$$

5B.2 Nonlinear Function of Two Parameters[17]

When working with functions of two (or more) parameters the approach is much the same, but the Taylor series approximation changes to a more general form. For a function of two parameters, the Taylor series approximation is

$$g_2(b_1, b_2) \cong g_2(\beta_1, \beta_2) + \frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}(b_1 - \beta_1) + \frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}(b_2 - \beta_2) \tag{5B.3}$$

1. If $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$, then $E[g_2(b_1, b_2)] \cong g_2(\beta_1, \beta_2)$.
2. If b_1 and b_2 are consistent estimators, so that $b_1 \xrightarrow{p} \beta_1$ and $b_2 \xrightarrow{p} \beta_2$, then $g_2(b_1, b_2) \xrightarrow{p} g_2(\beta_1, \beta_2)$.
3. The variance of $g_2(b_1, b_2)$ is given by the **delta method** as

$$\text{var}[g_2(b_1, b_2)] \cong \left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}\right]^2 \text{var}(b_1) + \left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}\right]^2 \text{var}(b_2)$$
$$+ 2\left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}\right]\left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}\right] \text{cov}(b_1, b_2) \tag{5B.4}$$

4. The estimator $g_2(b_1, b_2)$ has an approximate normal distribution in large samples,

$$g_2(b_1, b_2) \overset{a}{\sim} N(g_2(\beta_1, \beta_2),\ \text{var}[g_2(b_1, b_2)]) \tag{5B.5}$$

The asymptotic normality of $g_2(b_1, b_2)$ means that we can test nonlinear hypotheses such as $H_0 : g_2(\beta_1, \beta_2) = c$, and we can construct interval estimates of $g_2(\beta_1, \beta_2)$ in the usual way.

[17] This section contains advanced material. The general case involving a function of more than two parameters requires matrix algebra. See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorems D.21A and D.22 in online Appendix available at pages.stern.nyu.edu/~wgreene/text/econometricanalysis.htm.

In practice we evaluate the derivatives at the estimates b_1 and b_2, and the variances and covariances by their usual estimates from equations such as those for the simple regression model in (2.20)–(2.22).

EXAMPLE 5.20 | An Interval Estimate for β_1/β_2

The nonlinear function of two parameters that we consider is $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. To employ the delta method, we require the derivatives (see Appendix A, Derivative Rules 3 and 6)

$$\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1} = \frac{1}{\beta_2} \quad \text{and} \quad \frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2} = -\frac{\beta_1}{\beta_2^2}$$

The estimate $g_2(b_1, b_2) = b_1/b_2 = 87.44311/10.68456 = 8.18406$ and its estimated variance is

$$\begin{aligned}\widehat{\text{var}}[g_2(b_1, b_2)] &= \left[\frac{1}{b_2}\right]^2 \widehat{\text{var}}(b_1) + \left[-\frac{b_1}{b_2^2}\right]^2 \widehat{\text{var}}(b_2) \\ &\quad + 2\left[\frac{1}{b_2}\right]\left[-\frac{b_1}{b_2^2}\right] \widehat{\text{cov}}(b_1, b_2) \\ &= 22.61857\end{aligned}$$

The delta method standard error is $\text{se}(b_1/b_2) = 4.75590$. The resulting 95% interval estimate for β_1/β_2 is $(-1.807712, 18.17583)$. While all this seems incredibly complicated, most software packages will compute at least the estimates and standard errors automatically. And now that you understand the calculations, you can be confident when you use the "canned" routines.

Appendix 5C Monte Carlo Simulation

In Appendices 2H and 3C, we introduced a Monte Carlo simulation to illustrate the repeated sampling properties of the least squares estimators. In this appendix, we use the same framework to illustrate the repeated sampling performances of interval estimators and hypothesis tests when the errors are not normally distributed.

Recall that the **data generation process** for the simple linear regression model is given by

$$y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

The Monte Carlo parameter values are $\beta_1 = 100$ and $\beta_2 = 10$. The value of x_i is 10 for the first $N/2$ observations and 20 for the remaining $N/2$ observations, so that the regression functions are

$$\begin{aligned}E(y_i|x_i = 10) &= 100 + 10x_i = 100 + 10 \times 10 = 200, \quad i = 1, \ldots, N/2 \\ E(y_i|x_i = 20) &= 100 + 10x_i = 100 + 10 \times 20 = 300, \quad i = (N/2) + 1, \ldots, N\end{aligned}$$

5C.1 Least Squares Estimation with Chi-Square Errors

In this appendix, we modify the simulation in an important way. The random errors are independently distributed but with normalized chi-square distributions. In Figure B.7, the *pdf*s of several chi-square distributions are shown. We will use the $\chi^2_{(4)}$ in this simulation, which is skewed with a long tail to the right. Let $v_i \sim \chi^2_{(4)}$. The expected value and variance of this random variable are $E(v_i) = 4$ and $\text{var}(v_i) = 8$, respectively, so that $z_i = (v_i - 4)/\sqrt{8}$ has mean zero and variance one. The random errors we employ are $e_i = 50z_i$ so that $\text{var}(e_i|x_i) = \sigma^2 = 2500$, as in earlier appendices.

As before, we use $M = 10{,}000$ Monte Carlo simulations, using the sample sizes $N = 20$, 40 (as before), 100, 200, 500, and 1000. Our objectives are to illustrate that the least squares

estimators of β_1, β_2, and the estimator $\hat{\sigma}^2$ are unbiased, and to investigate whether hypothesis tests and interval estimates perform as they should, even though the errors are not normally distributed. As in Appendix 3C, we

- Test the null hypothesis $H_0 : \beta_2 = 10$ using the one-tail alternative $H_0 : \beta_2 > 10$. The critical value for the test is the 95th percentile of the t-distribution with $N - 2$ degrees of freedom, $t_{(0.95,\, N-2)}$. We report the percentage of rejections from this test (*REJECT*).
- Contruct a 95% interval estimate for β_2 and report the percentage of the estimates (*COVER*) that contain the true parameter, $\beta_2 = 10$.
- Compute the percentage of the time (*CLOSE*) that the estimates b_2 are in the interval $\beta_2 \pm 1$, or between 9 and 11. Based on our theory, this percentage should increase toward 1 as N increases.

The Monte Carlo simulation results are summarized in Table 5C.1.

The unbiasedness of the least squares estimators is verified by the average values of the estimates being very close to the true parameter values for all sample sizes. The percentage of estimates that are "close" to the true parameter value rises as the sample size N increases, verifying the consistency of the estimator. Because the rejection rates from the t-test are close to 0.05 and the coverage of the interval estimates is close to 95%, the approximate normality of the estimators is very good. To illustrate, in Figure 5C.1 we present the histogram of the estimates b_2 for $N = 40$.

TABLE 5C.1 **The Least Squares Estimators, Tests, and Interval Estimators**

N	$\bar{b}_1$	$\bar{b}_2$	$\overline{\hat{\sigma}^2}$	*REJECT*	*COVER*	*CLOSE*
20	99.4368	10.03317	2496.942	0.0512	0.9538	0.3505
40	100.0529	9.99295	2498.030	0.0524	0.9494	0.4824
100	99.7237	10.01928	2500.563	0.0518	0.9507	0.6890
200	99.8427	10.00905	2497.473	0.0521	0.9496	0.8442
500	100.0445	9.99649	2499.559	0.0464	0.9484	0.9746
1000	100.0237	9.99730	2498.028	0.0517	0.9465	0.9980

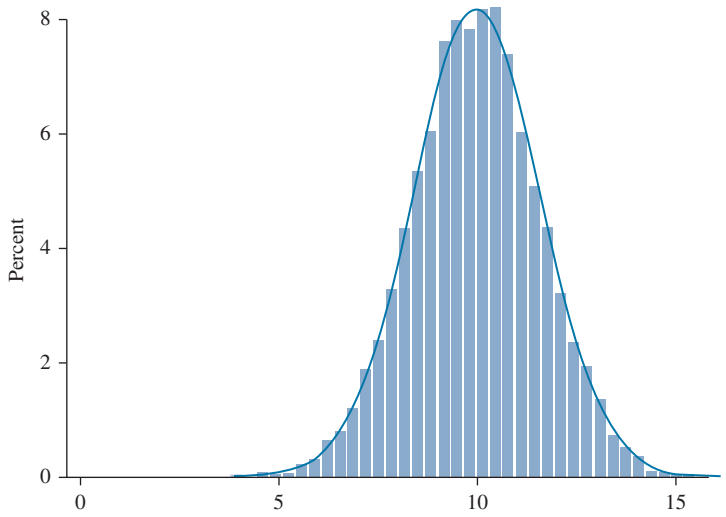


FIGURE 5C.1 **Histogram of the estimates b_2 for $N = 40$.**

It is very bell shaped, with the superimposed normal density function fitting it very well. The nonnormality of the errors does not invalidate inferences in this model, even with only $N = 40$ sample observations.

5C.2 Monte Carlo Simulation of the Delta Method

In this Monte Carlo simulation, again using 10,000 samples, we compute the value of the nonlinear function estimator $g_1(b_2) = \exp(b_2/10)$ for each sample, and we test the true null hypothesis $H_0 : g_1(\beta_2) = \exp(\beta_2/10) = e^1 = 2.71828$ using a two-tail test at the 5% level of significance. We are interested in how well the estimator does in finite samples (recall that the random errors are not normally distributed and that the function is nonlinear), and how well the test performs. In Table 5C.2, we report the average of the parameter estimates for each sample size. Note that the mean estimate converges toward the true value as N becomes larger. The test at the 5% level of significance rejects the true null hypothesis about 5% of the time. The test statistic is

$$t = \frac{g_1(b_2) - 2.71828}{\text{se}[g_1(b_2)]} \sim t_{(N-2)}$$

The fact that the t-test rejects the correct percentage of the time implies not only that the estimates are well behaved but also that the standard error in the denominator is correct, and that the distribution of the statistic is "close" to its limiting standard normal distribution. In Table 5C.2, $\overline{\text{se}[\exp(b_2/10)]}$ is the average of the nominal standard errors calculated using the delta method, and std. dev.$[\exp(b_2/10)]$ is the standard deviation of the estimates that measures the actual, true variation in the Monte Carlo estimates. We see that for sample sizes $N = 20$ and $N = 40$, the average of the standard errors calculated using the delta method is smaller than the true standard deviation, meaning that on average, in this illustration, the delta method overstates the precision of the estimates $\exp(b_2/10)$. The average standard error calculated using the delta method is close to the true standard deviation for larger sample sizes. We are reminded that the delta method standard errors are valid in large samples, and in this illustration the sample size $N = 100$ seems adequate for the asymptotic result to hold. The histogram of the estimates for sample size $N = 40$ in Figure 5C.2 shows only the very slightest deviation from normality, which is why the t-test performs so well.

We now examine how well the delta method works at different sample sizes for estimating the function $g_2(\beta_1/\beta_2)$ and approximating its variance and asymptotic distribution. The mean estimates in Table 5C.3 show that there is some bias in the estimates for small samples sizes. However, the bias diminishes as the sample size increases and is close to the true value, 10, when $N = 100$. The average of the delta method standard errors, $\overline{\text{se}(b_1/b_2)}$, is smaller than the actual, Monte Carlo, standard deviation of the estimates b_1/b_2 for sample sizes $N = 20$, 40, and 100. This illustrates the lesson that the more complicated the nonlinear function, or model, the larger the sample size that is required for asymptotic results to hold.

TABLE 5C.2 Simulation Results for $g_1(\beta_2) = \exp(\beta_2/10)$

N	$\overline{\exp(b_2/10)}$	$\overline{\text{se}[\exp(b_2/10)]}$	Std. dev. $[\exp(b_2/10)]$	*REJECT*
20	2.79647	0.60738	0.63273	0.0556
40	2.75107	0.42828	0.44085	0.0541
100	2.73708	0.27208	0.27318	0.0485
200	2.72753	0.19219	0.19288	0.0503
500	2.72001	0.12148	0.12091	0.0522
1000	2.71894	0.08589	0.08712	0.0555

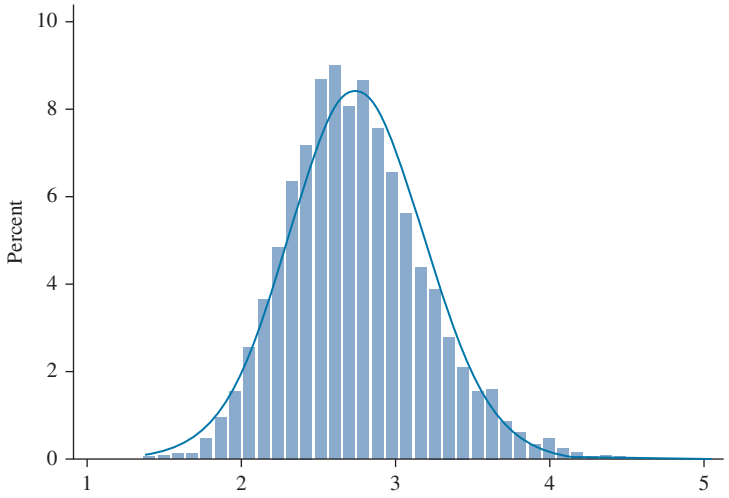


FIGURE 5C.2 Histogram of $g_1(b_2) = \exp(b_2/10)$.

TABLE 5C.3 Simulation Results for $g_2(b_1, b_2) = b_1/b_2$

N	$\overline{b_1/b_2}$	$\overline{\text{se}(b_1/b_2)}$	Std. dev. (b_1/b_2)
20	11.50533	7.18223	9.19427
40	10.71856	4.36064	4.71281
100	10.20997	2.60753	2.66815
200	10.10097	1.82085	1.82909
500	10.05755	1.14635	1.14123
1000	10.03070	0.80829	0.81664

The Monte Carlo simulated values of $g_2(b_1, b_2) = b_1/b_2$ are shown in Figures 5C.3(a) and (b) from the experiments with $N = 40$ and $N = 200$. With sample size $N = 40$, there is pronounced skewness. With $N = 200$, the distribution of the estimates is much more symmetric and bell shaped.

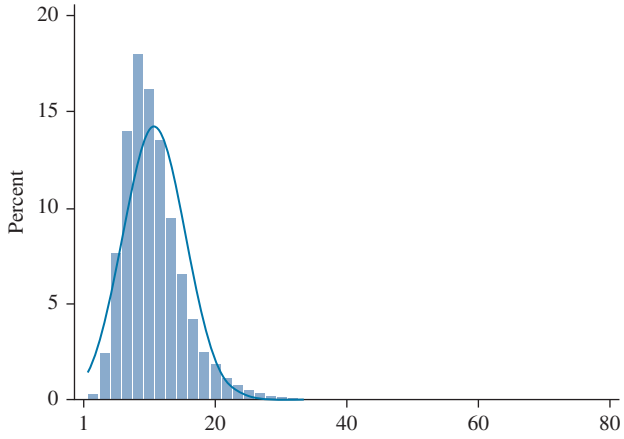


FIGURE 5C.3a Histogram of $g_2(b_1, b_2) = b_1/b_2$, $N = 40$.

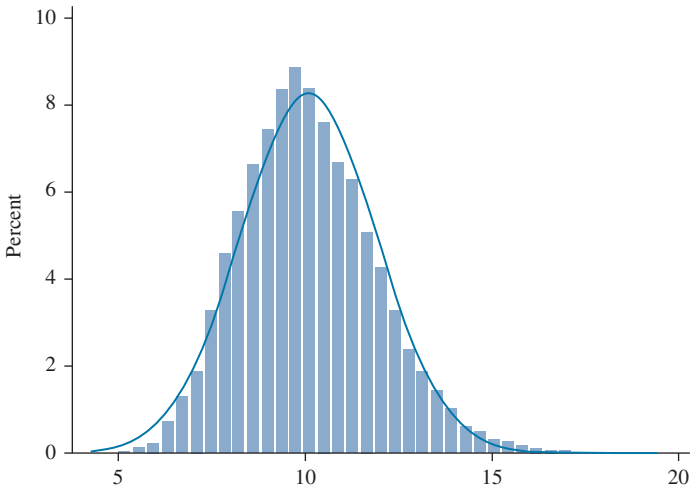


FIGURE 5C.3b **Histogram of $g_2(b_1, b_2) = b_1/b_2$, $N = 200$.**

Appendix 5D Bootstrapping

In Section 2.7.3, we discuss the interpretation of **standard errors** of estimators. Least squares estimates vary from sample to sample simply because the composition of the sample changes. This is called **sampling variability**. For the least squares estimators we have derived formulas for the variance of the least squares estimators. For example, in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the variance of the least squares estimator of the slope is $\text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum(x_i - \bar{x})^2$ and the standard error is $\text{se}(b_2) = \left[\hat{\sigma}^2 / \sum(x_i - \bar{x})^2\right]^{1/2}$. We were able to derive this formula using the model assumptions and linear form of the least squares estimator.

However, there are estimators for whom no easy standard errors can be computed. The estimators may be based on complex multistep procedures, or they may be nonlinear functions. In many cases, we can show that the estimators are **consistent** and **asymptotically normal**. We discussed these properties in Section 5.7. For an estimator $\hat{\beta}$, these properties mean that $\hat{\beta} \overset{a}{\sim} N\left[\beta, \text{var}\left(\hat{\beta}\right)\right]$. In this expression, $\text{var}\left(\hat{\beta}\right)$ is an **asymptotic variance** that is appropriate in large samples. If the asymptotic variance is known, then the **nominal standard error**, that is valid in large samples, is $\text{se}\left(\hat{\beta}\right) = \left[\widehat{\text{var}}\left(\hat{\beta}\right)\right]^{1/2}$. Asymptotic variance formulas can be difficult to derive. We illustrated the **delta method**, in Appendices 5B and 5C.2, for finding asymptotic variances of nonlinear functions of the least squares estimators. Even in those simple cases, there are derivatives and tedious algebra.

The **bootstrap procedure** is an alternative and/or complement to the analytic derivation of asymptotic variances. **Bootstrapping** can be used to compute standard errors for complicated and nonlinear estimators. It uses the speed of modern computing and a technique called **resampling**. In this section, we explain the bootstrapping technique and several ways that it can be used. In particular, we can use bootstrapping to

1. Estimate the bias of the estimator $\hat{\beta}$.
2. Obtain a standard error $\text{se}\left(\hat{\beta}\right)$ that is valid in large samples.
3. Construct confidence intervals for β.
4. Find critical values for test statistics.

5D.1 Resampling

To illustrate resampling suppose we have N independent and identically distributed data pairs (y_i, x_i). This is the case if we collect random samples from a specific population.[18] To keep matters simple let $N = 5$. This is for illustration only. A hypothetical sample is given in Table 5D.1. Resampling means randomly select $N = 5$ rows **with replacement** to form a new sample. The phrase **with replacement** means that after randomly selecting one row, and adding it to a new data set, we return the selected row to the original data where it might be randomly selected again, or not.

Perhaps seeing an algorithm for doing this will help. It begins with the concept of a **uniform random number** on the zero to one interval, $u \sim \text{uniform}(0,1)$. Uniform random numbers are a core part of numerical methods for simulations. We discuss them in Appendix B.4.1. Roughly speaking, the uniformly distributed random value u is equally likely to take any value in the interval (0,1). Computer scientists have designed algorithms so that repeated draws using a **uniform random number generator** are independent of one another. These are built into every econometric software package, although the algorithms used may vary slightly from one to the next. To randomly pick a row of data,

1. Let $u^* = (5 \times u) + 1$. This value is greater than 1 but less than 6.
2. Drop the decimal portion to obtain a random integer b that is 1, 2, 3, 4, or 5.

Table 5D.2 illustrates the process for $N = 5$. These steps are automated by many software packages, so you will not have to do the programming yourself, but it is a good idea to know what is happening. The values j in Table 5D.2 are the rows from the original data set that will constitute the first **bootstrap sample**. The first bootstrap sample will contain observations 5, 1, 2, and the third observation twice, as shown in Table 5D.3.[19] This is perfectly OK. Resampling means that

TABLE 5D.1 The Sample

Observation	y	x
1	$y_1 = 6$	$x_1 = 0$
2	$y_2 = 2$	$x_2 = 1$
3	$y_3 = 3$	$x_3 = 2$
4	$y_4 = 1$	$x_4 = 3$
5	$y_5 = 0$	$x_5 = 4$

TABLE 5D.2 Random Integers

u	u^*	j
0.9120440	5.56022	5
0.0075452	1.037726	1
0.2808588	2.404294	2
0.4602787	3.301394	3
0.5601059	3.800529	3

[18] Bootstrap techniques for time-series data are much different, and we will not discuss them here.

[19] Random number generators use a "starting value," called a **seed**. By choosing a seed the same sequence of random numbers can be obtained in subsequent runs. See Appendix B.4.1 for a discussion of how one class of random number generators work.

TABLE 5D.3 One Bootstrap Sample

Observation	y	x
5	$y_5 = 0$	$x_5 = 4$
1	$y_1 = 6$	$x_1 = 0$
2	$y_2 = 2$	$x_2 = 1$
3	$y_3 = 3$	$x_3 = 2$
3	$y_3 = 3$	$x_3 = 2$

some observations will be chosen multiple times, and others (such as observation 4 in this case) will not appear at all.

5D.2 Bootstrap Bias Estimate

The estimator $\hat{\beta}$ may be a biased estimator. Estimator bias is the difference between the estimator's expected value and the true parameter, or

$$\text{bias}\left(\hat{\beta}\right) = E\left(\hat{\beta}\right) - \beta$$

For a consistent estimator the bias disappears as $N \to \infty$, but we can estimate the bias given a sample of size N. Using the process described in the previous section, obtain bootstrap samples $b = 1, 2, \ldots, B$, each of size N. Using each bootstrap sample obtain an estimate $\hat{\beta}_b$. If $B = 200$, then we have 200 bootstrap sample estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{200}$. The average, or sample mean, of the B bootstrap sample estimates is

$$\bar{\hat{\beta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b$$

The bootstrap estimate of the bias is

$$\text{bootstrap } \widehat{\text{bias}}\left(\hat{\beta}\right) = \bar{\hat{\beta}} - \hat{\beta}_O$$

where $\hat{\beta}_O$ is the estimate obtained using the original sample [the subscript is "oh" and not zero]. In this calculation, $\bar{\hat{\beta}}$ plays the role of $E\left(\hat{\beta}\right)$ and $\hat{\beta}_O$, the estimate from the original sample, plays the role of the true parameter β. A descriptive saying about bootstrapping is that that "$\hat{\beta}_O$ is true in the sample," emphasizing the role played by the original sample estimate, $\hat{\beta}_O$.

5D.3 Bootstrap Standard Error

Bootstrap standard error calculation requires B bootstrap samples of size N. For the purpose of computing standard errors, the number of bootstrap samples should be at least 50, and perhaps 200 or 400, depending on the complexity of your estimation problem.[20] The bootstrap standard error is the **sample standard deviation** of the B bootstrap estimates. The sample standard deviation is the square root of the sample variance. The bootstrap estimate of $\text{var}\left(\hat{\beta}\right)$ is the sample variance of the bootstrap estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_B$,

$$\text{bootstrap var}\left(\hat{\beta}\right) = \sum_{b=1}^{B}\left(\hat{\beta}_b - \bar{\hat{\beta}}\right)^2/(B-1)$$

[20] Try a number of bootstraps B. For standard errors $B = 200$ is a good starting value. Compute the bootstrap standard error. Change the random number seed a few times. If the bootstrap standard error changes little, then B is large enough. If there are substantial changes, increase B.

The bootstrap standard error is

$$\text{bootstrap se}\left(\hat{\beta}\right) = \sqrt{\text{bootstrap var}\left(\hat{\beta}\right)} = \sqrt{\sum_{b=1}^{B}\left(\hat{\beta}_b - \bar{\hat{\beta}}\right)^2/(B-1)}$$

In large samples, the bootstrap standard error is no better, or worse, than the theoretically derived standard error. The advantage of the bootstrap standard error is that we need not derive the theoretical standard error, which can sometimes be very difficult. Even if the theoretical standard error can be obtained, the bootstrap standard error can be used as a check of the estimate based on a theoretical formula. If the bootstrap standard error is considerably different from the theory-based standard error, then either (i) the sample size N is not large enough to justify asymptotic theory, or (ii) the theoretical formula has an error. The theoretical standard error could be wrong if one of the model assumptions does not hold, or there is a math error, or there is an error in the software calculating the estimate based on the theoretical standard error (yes, that sometimes happens).

We can use the bootstrap standard error the same way as the usual standard error. An asymptotically justified $100(1-\alpha)\%$ interval estimator of β is

$$\hat{\beta} \pm t_c\left[\text{bootstrap se}\left(\hat{\beta}\right)\right]$$

where t_c is the $1-\alpha/2$ percentile of the t-distribution. In large samples, using $t_c = 1.96$ leads to a 95% interval estimate. This is sometimes called the **normal-based bootstrap confidence interval**.

For testing the null hypothesis $H_0: \beta = c$ against $H_1: \beta \neq c$, a valid test statistic is

$$t = \frac{\hat{\beta} - c}{\text{bootstrap se}\left(\hat{\beta}\right)}$$

If the null hypothesis is true, the test statistic has a standard normal distribution[21] in large samples. At the 5% level, we reject the null hypothesis if $t \geq 1.96$ or $t \leq -1.96$.

5D.4 Bootstrap Percentile Interval Estimate

A **percentile interval estimate**, or **percentile confidence interval**, does not use the approximate large sample normality of an estimator. Recall that in the simple regression model a 95% interval estimator is obtained from equation (3.5), which is

$$P\left[b_k - t_c\text{se}\left(b_k\right) \leq \beta_k \leq b_k + t_c\text{se}\left(b_k\right)\right] = 1-\alpha$$

where $t_c = t_{(0.975,\,N-K)}$. The interval estimator $\left[b_k - t_c\text{se}\left(b_k\right), b_k + t_c\text{se}\left(b_k\right)\right]$ will contain the true parameter β_k in 95% of **repeated samples** from the same population. Another descriptive phrase used when discussing bootstrapping is that we "treat the sample as the population." This makes the point that by using bootstrapping, we are trying to learn about an estimator's **sampling properties**; or how the estimator performs in repeated samples. Bootstrapping treats each bootstrap sample as a "repeated sample." Using this logic, if we obtain many bootstrap samples, and many estimates (sorting the B bootstrap estimates from smallest to largest) a 95% percentile interval estimate is $\left[\hat{\beta}^*_{(0.025)},\ \hat{\beta}^*_{(0.975)}\right]$ where $\hat{\beta}^*_{(0.025)}$ is the 2.5%-percentile of the B bootstrap estimates, and $\hat{\beta}^*_{(0.975)}$ is the 97.5%-percentile of the B bootstrap estimates. Because of the way software programmers find percentiles, it is useful to choose B such that $\alpha(B+1)$ is a convenient integer. If $B = 999$, then the 2.5%-percentile is the 25th value and the 97.5%-percentile is the 975th value. If $B = 1999$, then the 2.5%-percentile is the 50th value and the 97.5%-percentile is the 1950th value. Calculating percentile interval estimates requires a larger number of bootstrap samples than calculating a standard error. Intervals calculated this way are not necessarily symmetrical.

[21] Because of its large sample justification, some software packages will call this statistic "z."

5D.5 Asymptotic Refinement

If it is possible to derive a theoretical expression for the variance of an estimator that is valid in large samples, then we can combine it with bootstrapping to improve upon standard asymptotic theory. Asymptotic refinement produces a test statistic critical value that leads to more accurate tests. What do we mean by that? A test of $H_0: \beta = c$ against $H_1: \beta \neq c$ uses an asymptotically valid nominal standard error and the t-statistic $t = \left(\hat{\beta} - c\right)/\text{se}\left(\hat{\beta}\right)$. If $\alpha = 0.05$, we reject the null hypothesis if $t \geq 1.96$ or $t \leq -1.96$. This test is called a **symmetrical two-tail test**. In finite (small) samples, the actual rejection probability is not $\alpha = 0.05$ but $P(\text{reject } H_0 | H_0 \text{ is } \textit{true}) = \alpha + \textit{error}$. The *error* goes to zero as the sample size N approaches infinity. More precisely, $N \times \textit{error} \leq N^*$ where N^* is some upper bound. In order for this to be true, as $N \to \infty$ the *error* must approach zero, $\textit{error} \to 0$. Not only must $\textit{error} \to 0$, but also it must approach zero at the same rate as $N \to \infty$, so that the two effects are offsetting, with product $N \times \textit{error}$ staying a finite number. This is called convergence to zero at rate "N." Using a bootstrap critical value, t_c^*, instead of 1.96 it can be shown that $N^2 \times \textit{error} \leq N^*$, so that the test size *error* converges to zero at rate N^2. We have a more accurate test because the *error* in the test size goes to zero faster using the bootstrap critical value.

The gain in accuracy is "easy" to obtain. Resample the data B times. In each bootstrap sample, compute

$$t_b = \frac{\hat{\beta}_b - \hat{\beta}_O}{\text{se}\left(\hat{\beta}_b\right)}$$

In this expression, $\hat{\beta}_b$ is the estimate in the bth bootstrap sample, $\hat{\beta}_O$ is the estimate based on the original sample, and $\text{se}\left(\hat{\beta}_b\right)$ is the nominal standard error, the usual theory-based standard error, calculated using the bth bootstrap sample. This is the bootstrap equivalent of equation (3.3). To find the bootstrap critical value t_c^* (i) compute $|t_b|$, (ii) sort them in ascending magnitude, then (iii) t_c^* is the $100(1-\alpha)$-percentile of $|t_b|$. To test $H_0: \beta = c$ against $H_1: \beta \neq c$ use the t-statistic $t = \left(\hat{\beta} - c\right)/\text{se}\left(\hat{\beta}\right)$ computed with the original sample, and reject the null hypothesis if $t \geq t_c^*$ or $t \leq -t_c^*$. The $100(1-\alpha)\%$ interval estimate $\hat{\beta} \pm t_c^* \text{se}\left(\hat{\beta}\right)$ is sometimes called a **percentile-*t*** interval estimate.

For a right-tail test, $H_0: \beta \leq c$ against $H_1: \beta > c$, t_c^* is the $100(1-\alpha)$-percentile of t_b, dropping the absolute value operation. Reject the null hypothesis if $t \geq t_c^*$. For a left-tail test, $H_0: \beta \geq c$ against $H_1: \beta < c$, t_c^* is the 100α-percentile of t_b. Reject the null hypothesis if $t \leq t_c^*$.

EXAMPLE 5.21 | Bootstrapping for Nonlinear Functions $g_1(\beta_2) = \exp(\beta_2/10)$ and $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$.

Clearly it is time for an example! Using the same Monte Carlo design as in Appendix 5C, we create one sample for $N = 20$, 40, 100, 200, 500, and 1000. They are in the data files *mc20*, *mc40*, *mc100*, *mc200*, *mc500*, and *mc1000*.

First we explore bootstrapping $g_1(\beta_2) = \exp(\beta_2/10)$. Table 5D.4a contains the estimates, delta method standard error, and an asymptotically justified 95% interval estimate

$$\exp(b_2/10) \pm \left\{1.96 \times \text{se}\left[\exp(b_2/10)\right]\right\}$$

Compare these to Table 5C.2 containing the Monte Carlo *averages* of the estimates, the nominal (delta method) standard errors, and the standard deviation of the estimates.

Because we will calculate percentile interval estimates and a bootstrap critical value, we use $B = 1999$ bootstrap samples as the basis for the estimates in Table 5D.4b. The bootstrap estimates of the bias diminish as the sample size increases, reflecting the consistency of the estimator. The bootstrap standard errors for $N = 20$, 40, and 100 are quite similar to the delta method standard errors for these sample sizes shown in Table 5D.4a. They are not as similar to the Monte Carlo average nominal standard error and standard deviation in Table 5C.2. However, once the sample size is $N = 200$ or more, the bootstrap standard errors are much closer to the results in Table 5C.2. In Table 5D.4b, we also

TABLE 5D.4a Delta Method $g_1(\beta_2) = \exp(\beta_2/10) = 2.71828$

N	$g_1(b_2) = \exp(b_2/10)$	se$[\exp(b_2/10)]$	95% Interval
20	2.91088	0.62367	[1.6885, 4.1332]
40	2.34835	0.37781	[1.6079, 3.0888]
100	2.98826	0.30302	[2.3945, 3.5822]
200	2.86925	0.20542	[2.4666, 3.2719]
500	2.63223	0.11241	[2.4119, 2.8526]
1000	2.78455	0.08422	[2.6195, 2.9496]

TABLE 5D.4b Bootstrapping $g_1(\beta_2) = \exp(\beta_2/10)$

N	Bootstrap Bias	Bootstrap se	PI	t_c^*
20	0.0683	0.6516	[2.0098, 4.5042]	3.0063
40	0.0271	0.3796	[1.7346, 3.2173]	2.2236
100	0.0091	0.3050	[2.4092, 3.6212]	2.0522
200	0.0120	0.2039	[2.4972, 3.3073]	1.9316
500	−0.0001	0.1130	[2.4080, 2.8567]	2.0161
1000	0.0025	0.0844	[2.6233, 2.9593]	1.9577

report the 95% **percentile interval** (PI) **estimate** for each sample size. Finally, we report the asymptotically refined critical value that would be used for a symmetrical two-tail test at the 5% level of significance, or when constructing a confidence interval. Based on these values, we judge that sample sizes $N = 20$ and 40 are not really sufficiently large to support asymptotic inferences in our specific samples, but if we do proceed, then the usual critical value 1.96 should not be used for t-tests or interval estimates. For sample sizes $N = 100$ or more, it appears that usual asymptotic procedures can be justified.

Table 5D.5 contains similar results for the function $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. The estimates, bootstrap bias, delta method standard error, and bootstrap standard error tell a similar story. For this nonlinear function, a ratio of two parameters, $N = 200$ or more would make us feel better about asymptotic inference. It is reassuring when the bootstrap and delta method standard errors are similar, although these are somewhat smaller than the average nominal standard error and standard deviations in Table 5C.3. Expressions containing ratios of parameters in one form or another often require larger samples for asymptotic inference to hold.

TABLE 5D.5 Bootstrapping $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$

N	$g_2(b_1, b_2) = b_1/b_2$	Bootstrap Bias	se(b_1/b_2)	Bootstrap se
20	8.18406	0.7932	4.75590	4.4423
40	13.15905	1.0588	5.38959	6.0370
100	7.59037	0.2652	2.14324	2.3664
200	8.71779	0.0714	1.64641	1.6624
500	10.74195	0.0825	1.15712	1.2180
1000	9.44545	0.0120	0.73691	0.7412

CHAPTER 6

Further Inference in the Multiple Regression Model

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the concepts of restricted and unrestricted sums of squared errors and how they are used to test hypotheses.
2. Use the F-test to test single null hypotheses or joint null hypotheses.
3. Use your computer software to perform an F-test.
4. Test the overall significance of a regression model and identify the components of this test from your computer output.
5. From output of your computer software, locate (a) the sum of squared errors, (b) the F-value for the overall significance of a regression model, (c) the estimated covariance matrix for the least squares estimates, and (d) the correlation matrix for the explanatory variables.
6. Explain the relationship between the finite sample F-test and the large sample χ^2-test, and the assumptions under which each is suitable.
7. Obtain restricted least squares estimates that include nonsample information in the estimation procedure.
8. Explain the properties of the restricted least squares estimator. In particular, how do its bias and variance compare with those of the unrestricted, ordinary, least squares estimator?
9. Explain the differences between models designed for prediction and models designed to estimate a causal effect.
10. Explain what is meant by (a) an omitted variable and (b) an irrelevant variable. Explain the consequences of omitted and irrelevant variables for the properties of the least squares estimator.
11. Explain the concept of a control variable and the assumption necessary for a control variable to be effective.
12. Explain the issues that need to be considered when choosing a regression model.
13. Test for misspecification using RESET.
14. Compute forecasts, standard errors of forecast errors, and interval forecasts from a multiple regression model.
15. Use the Akaike information or Schwartz criteria to select variables for a predictive model.

16. Identify collinearity and explain its consequences for least squares estimation.
17. Identify influential observations in a multiple regression model.
18. Compute parameter estimates for a regression model that is nonlinear in the parameters and explain how nonlinear least squares differs from linear least squares.

KEYWORDS

χ^2-test
AIC
auxiliary regressions
BIC
causal model
collinearity
control variables
F-test
influential observations
irrelevant variables
nonlinear least squares
nonsample information
omitted variable bias
overall significance
prediction
predictive model
RESET
restricted least squares
restricted model
restricted *SSE*
SC
single and joint null hypotheses
unrestricted model
unrestricted *SSE*

Economists develop and evaluate theories about economic behavior. Hypothesis testing procedures are used to test these theories. In Chapter 5, we developed *t*-tests for null hypotheses consisting of a single restriction on one parameter β_k from the multiple regression model, and null hypotheses consisting of a single restriction that involves more than one parameter. In this chapter we extend our earlier analysis to testing a null hypothesis with two or more restrictions on two or more parameters. An important new development for such tests is the *F*-test. A large sample alternative that can be used under weaker assumptions is the χ^2-test.

The theories that economists develop sometimes provide nonsample information that can be used along with the information in a sample of data to estimate the parameters of a regression model. A procedure that combines these two types of information is called restricted least squares. It can be a useful technique when the data are not information-rich—a condition called collinearity—and the theoretical information is good. The restricted least squares procedure also plays a useful practical role when testing hypotheses. In addition to these topics, we discuss model specification for the multiple regression model, prediction, and the construction of prediction intervals. Model specification involves choosing a functional form and choosing a set of explanatory variables.

Critical to the choice of a set of explanatory variables is whether a model is to be used for prediction or causal analysis. For causal analysis, omitted variable bias and selection of control variables is important. For prediction, selection of variables that are highly correlated with the dependent variable is more relevant. We also discuss the problems that arise if our data are not sufficiently rich because the variables are collinear or lack adequate variation, and summarize concepts for detecting influential observations. The use of nonlinear least squares is introduced for models that are nonlinear in the parameters.

6.1 Testing Joint Hypotheses: The *F*-test

In Chapter 5 we showed how to use one- and two-tail *t*-tests to test hypotheses involving

1. A single coefficient
2. A linear combination of coefficients
3. A nonlinear combination of coefficients.

The test for a single coefficient was the most straightforward, requiring only the estimate of the coefficient and its standard error. For testing a linear combination of coefficients, computing the standard error of the estimated linear combination brought added complexity. It uses the variances and covariances of all estimates in the linear combination and can be computationally demanding if done on a hand calculator, especially if there are three or more coefficients in the linear combination. Software will perform the test automatically, however, yielding the standard error, the value of the t-statistic, and the p-value of the test. If assumptions MR1–MR6 hold then t-statistics have exact distributions, making the tests valid for small samples. If MR6 is violated, implying $(e_i|\mathbf{X})$ is no longer normally distributed, or if MR2: $E(e_i|\mathbf{X}) = 0$ is weakened to the conditions $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0$, then we need to rely on large sample results that make the tests approximately valid, with the approximation improving as sample size increases.

For testing non-linear combinations of coefficients, one must rely on large sample approximations even if assumptions MR1–MR6 hold, and the delta method must be used to compute standard errors. Derivatives of the nonlinear function and the covariance matrix of the coefficients are required, but as with a linear combination, software will perform the test automatically, computing the standard error for you, as well as the value of the t-statistic and its p-value. In Chapter 5 we gave an example of an interval estimate rather than a hypothesis test for a nonlinear combination, but that example—the optimal level of advertising—showed how to obtain all the ingredients needed for a test. For both hypothesis testing and interval estimation of a nonlinear combination, it is the standard error that requires more effort.

A characteristic of all the t tests in Chapter 5 is that they involve a single conjecture about one or more of the parameters—or, put another way, there is only one "equal sign" in the null hypothesis. In this chapter, we are interested in extending hypothesis testing to null hypotheses that involve multiple conjectures about the parameters. A null hypothesis with multiple conjectures, expressed with more than one equal sign, is called a **joint hypothesis**. An example of a joint hypothesis is testing whether a group of explanatory variables should be included in a particular model. Should variables on socioeconomic background, along with variables describing education and experience, be used to explain a person's wage? Does the quantity demanded of a product depend on the prices of substitute goods, or only on its own price? Economic hypotheses such as these must be formulated into statements about model parameters. To answer the first of the two questions, we set up a null hypothesis where the coefficients of all the socioeconomic variables are equal to zero. For the second question, the null hypothesis would equate the coefficients of prices of all substitute goods to zero. Both are of the form

$$H_0 : \beta_4 = 0, \ \beta_5 = 0, \ \beta_6 = 0 \tag{6.1}$$

where β_4, β_5, and β_6 are the coefficients of the socioeconomic variables, or the coefficients of the prices of substitute goods. The joint null hypothesis in (6.1) contains three conjectures (three equal signs): $\beta_4 = 0$, $\beta_5 = 0$, and $\beta_6 = 0$. A test of H_0 is a joint test for whether all three conjectures hold simultaneously.

It is convenient to develop the test statistic for testing hypotheses such as (6.1) within the context of an example. We return to Big Andy's Burger Barn.

EXAMPLE 6.1 | Testing the Effect of Advertising

The test used for testing a joint null hypothesis is the ***F*-test**. To introduce this test and concepts related to it, consider the Burger Barn sales model given in (5.23):

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.2}$$

Suppose now we wish to test whether *SALES* is influenced by advertising. Since advertising appears in (6.2) as both a linear term *ADVERT* and as a quadratic term $ADVERT^2$, advertising will have no effect on sales if $\beta_3 = 0$ and $\beta_4 = 0$; advertising will have an effect if $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or if both β_3

and β_4 are nonzero. Thus, for this test our null and alternative hypotheses are

$$H_0 : \beta_3 = 0,\ \beta_4 = 0$$

$$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both are nonzero}$$

Relative to the null hypothesis $H_0 : \beta_3 = 0, \beta_4 = 0$, the model in (6.2) is called the **unrestricted model**; the restrictions in the null hypothesis have not been imposed on the model. It contrasts with the **restricted model**, which is obtained by assuming the parameter restrictions in H_0 are true. When H_0 is true, $\beta_3 = 0$ and $\beta_4 = 0$, and *ADVERT* and $ADVERT^2$ drop out of the model. It becomes

$$SALES = \beta_1 + \beta_2 PRICE + e \tag{6.3}$$

The *F*-test for the hypothesis $H_0 : \beta_3 = 0, \beta_4 = 0$ is based on a comparison of the sums of squared errors (sums of squared OLS residuals) from the unrestricted model in (6.2) and the restricted model in (6.3). Our shorthand notation for these two quantities is SSE_U and SSE_R, respectively.

Adding variables to a regression reduces the sum of squared errors—more of the variation in the dependent variable becomes attributable to the variables in the regression and less of its variation becomes attributable to the error. In terms of our notation, $SSE_R - SSE_U \geq 0$. Using the data in the file *andy* to estimate (6.2) and (6.3), we find that $SSE_U = 1532.084$ and $SSE_R = 1896.391$. Adding *ADVERT* and $ADVERT^2$ to the equation reduces the sum of squared errors from 1896.391 to 1532.084.

What the *F*-test does is to assess whether the reduction in the sum of squared errors is sufficiently large to be significant. If adding the extra variables has little effect on the sum of squared errors, then those variables contribute little to explaining variation in the dependent variable, and there is support for a null hypothesis that drops them. On the other hand, if adding the variables leads to a big reduction in the sum of squared errors, those variables contribute significantly to explaining the variation in the dependent variable, and we have evidence against the null hypothesis. The *F*-statistic determines what constitutes a large reduction or a small reduction in the sum of squared errors. It is given by

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \tag{6.4}$$

where J is the number of restrictions or number of hypotheses in H_0, N is the number of observations, and K is the number of coefficients in the unrestricted model.

To use the *F*-statistic to assess whether a reduction in the sum of squared errors is sufficient to reject the null hypothesis, we need to know its probability distribution when the null hypothesis is true. If assumptions MR1–MR6 hold, then, when the **null hypothesis is true**, the statistic F has what is called an *F*-distribution with J numerator degrees of freedom and $(N - K)$ denominator degrees of freedom. Some details about this distribution are given in Appendix B.3.8, with its typical shape illustrated in Figure B.9(a). **If the null hypothesis is not true,** then the difference between SSE_R and SSE_U becomes large, implying that the restrictions placed on the model by the null hypothesis significantly reduce the ability of the model to fit the data. A large value for $SSE_R - SSE_U$ means that the value of F tends to be *large*, so that we *reject* the null hypothesis if the value of the *F*-test statistic becomes too large. What is too large is decided by comparing the value of F to a critical value F_c, which leaves a probability α in the upper tail of the *F*-distribution with J and $N - K$ degrees of freedom. Tables of critical values for $\alpha = 0.01$ and $\alpha = 0.05$ are provided in Statistical Tables 4 and 5. The rejection region $F \geq F_c$ is illustrated in Figure B.9(a).

EXAMPLE 6.2 | The *F*-Test Procedure

Using the hypothesis testing steps introduced in Chapter 3, the *F*-test procedure for testing whether *ADVERT* and $ADVERT^2$ should be excluded from the sales equation is as follows:

1. *Specify the null and alternative hypotheses*: The joint null hypothesis is $H_0 : \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is $H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both are nonzero.

2. *Specify the test statistic and its distribution if the null hypothesis is true*: Having two restrictions in H_0 means $J = 2$. Also, recall that $N = 75$, so the distribution of the F-test statistic when H_0 is true is

$$F = \frac{(SSE_R - SSE_U)/2}{SSE_U/(75-4)} \sim F_{(2,71)}$$

3. *Set the significance level and determine the rejection region*: Using $\alpha = 0.05$, the critical value from the $F_{(2,71)}$-distribution is $F_c = F_{(0.95,2,71)}$, giving a rejection region of $F \geq 3.126$. Alternatively, H_0 is rejected if p-value ≤ 0.05.

4. *Calculate the sample value of the test statistic and, if desired, the p-value*: The value of the F-test statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(1896.391 - 1532.084)/2}{1532.084/(75-4)} = 8.44$$

The corresponding p-value is $p = P(F_{(2,71)} > 8.44) = 0.0005$.

5. *State your conclusion*: Since $F = 8.44 > F_c = 3.126$, we reject the null hypothesis that both $\beta_3 = 0$ and $\beta_4 = 0$, and conclude that at least one of them is not zero. Advertising does have a significant effect upon sales revenue. The same conclusion is reached by noting that p-value $= 0.0005 < 0.05$.

You might ask where the value $F_c = F_{(0.95,2,71)} = 3.126$ came from. The F critical values in Statistical Tables 4 and 5 are reported for only a limited number of degrees of freedom. However, exact critical values such as the one for this problem can be obtained for any number of degrees of freedom using your econometric software.

6.1.1 Testing the Significance of the Model

An important application of the F-test is for what is called testing the **overall significance** of a model. In Section 5.5.1, we tested whether the dependent variable y is related to a particular explanatory variable x_k using a t-test. In this section, we extend this idea to a joint test of the relevance of *all* the included explanatory variables. Consider again the general multiple regression model with $(K-1)$ explanatory variables and K unknown coefficients

$$y = \beta_1 + x_2\beta_2 + x_3\beta_3 + \cdots + x_K\beta_K + e \tag{6.5}$$

To examine whether we have a viable explanatory model, we set up the following null and alternative hypotheses:

$$H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$$
$$H_1 : \textit{At least one} \text{ of the } \beta_k \text{ is nonzero for } k = 2, 3, \ldots, K \tag{6.6}$$

The null hypothesis is a joint one because it has $K-1$ components. It conjectures that each and every one of the parameters β_k, other than the intercept parameter β_1, are simultaneously zero. If this null hypothesis is true, none of the explanatory variables influence y, and thus our model is of little or no value. If the alternative hypothesis H_1 is true, then at least one of the parameters is not zero, and thus one or more of the explanatory variables should be included in the model. The alternative hypothesis does not indicate, however, which variables those might be. Since we are testing whether or not we have a viable explanatory model, the test for (6.6) is sometimes referred to as a **test of the overall significance of the regression model**. Given that the t-distribution can only be used to test a single null hypothesis, we use the F-test for testing the joint null hypothesis in (6.6). The unrestricted model is that given in (6.5). The restricted model, assuming the null hypothesis is true, becomes

$$y_i = \beta_1 + e_i \tag{6.7}$$

The least squares estimator of β_1 in this restricted model is $b_1^* = \sum_{i=1}^{N} y_i/N = \bar{y}$, which is the sample mean of the observations on the dependent variable. The *restricted* sum of squared errors

from the hypothesis (6.6) is

$$SSE_R = \sum_{i=1}^{N}\left(y_i - b_1^*\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = SST$$

In this one case, in which we are testing the null hypothesis that all the model parameters are zero *except the intercept*, the restricted sum of squared errors is the total sum of squares (SST) from the full unconstrained model. The unrestricted sum of squared errors is the sum of squared errors from the unconstrained model—that is, $SSE_U = SSE$. The number of restrictions is $J = K - 1$. Thus, to test the overall significance of a model, *but not in general*, the F-test statistic can be modified and written as

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} \tag{6.8}$$

The calculated value of this test statistic is compared to a critical value from the $F_{(K-1,\,N-K)}$ distribution. It is used to test the overall significance of a regression model. The outcome of the test is of fundamental importance when carrying out a regression analysis, and it is usually automatically reported by computer software as the F-value.

EXAMPLE 6.3 | Overall Significance of Burger Barns Equation

To illustrate, we test the overall significance of the regression, (6.2), used to explain Big Andy's sales revenue. We want to test whether the coefficients of *PRICE*, *ADVERT*, and $ADVERT^2$ are all zero, against the alternative that at least one of these coefficients is not zero. Recalling that the model is $SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$, the hypothesis testing steps are as follows:

1. We are testing

$$H_0 : \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

against the alternative

$$H_1 : \text{At least one of } \beta_2 \text{ or } \beta_3 \text{ or } \beta_4 \text{ is nonzero}$$

2. If H_0 is true, $F = \dfrac{(SST - SSE)/(4-1)}{SSE/(75-4)} \sim F_{(3,71)}$.
3. Using a 5% significance level, we find the critical value for the F-statistic with (3,71) degrees of freedom is $F_c = 2.734$. Thus, we reject H_0 if $F \geq 2.734$.
4. The required sums of squares are $SST = 3115.482$ and $SSE = 1532.084$ which give an F-value of

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} = \frac{(3115.482 - 1532.084)/3}{1532.084/(75-4)} = 24.459$$

Also, p-value $= P(F \geq 24.459) = 0.0000$, correct to four decimal places.

5. Since $24.459 > 2.734$, we reject H_0 and conclude that the estimated relationship is a significant one. A similar conclusion is reached using the p-value. We conclude that at least one of *PRICE*, *ADVERT*, or $ADVERT^2$ have an influence on sales. Note that this conclusion is consistent with conclusions that would be reached using separate t-tests for the significance of each of the coefficients in (5.25).

Go back and check the output from your computer software. Can you find the F-value 24.459 and the corresponding p-value of 0.0000 that form part of the routine output?

6.1.2 The Relationship Between *t*- and *F*-Tests

A question that may have occurred to you is what happens if we have a null hypothesis which is not a joint hypothesis; it only has one equality in H_0? Can we use an F-test for this case, or do we go back and use a t-test? The answer is when testing a single "equality" null hypothesis (a single restriction) against a "not equal to" alternative hypothesis, either a t-test or an F-test can be used; the test outcomes will be identical. Two-tail t-tests are equivalent to F-tests *when there is*

a single hypothesis in H_0. An F-test cannot be used as an alternative to a one-tail t-test, however. To explore these notions we return to the Big Andy example.

EXAMPLE 6.4 | When are *t*- and *F*-tests equivalent?

In Examples 6.1 and 6.2, we tested whether advertising affects sales by using an F-test to test whether $\beta_3 = 0$ and $\beta_4 = 0$ in the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.9}$$

Suppose now we want to test whether *PRICE* affects *SALES*. Following the same F-testing procedure, we have $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$, and the restricted model

$$SALES = \beta_1 + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.10}$$

Estimating (6.9) and (6.10) gives $SSE_U = 1532.084$ and $SSE_R = 2683.411$, respectively. The required F-value is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(2683.411 - 1532.084)/1}{1532.084/(75-4)} = 53.355$$

The 5% critical vale is $F_c = F_{(0.95,\,1,\,71)} = 3.976$. Thus, we reject $H_0: \beta_2 = 0$.

Now let us see what happens if we use a t-test for the same problem: $H_0: \beta_2 = 0$ and $H_1: \beta_2 \neq 0$. The results from estimating (6.9) were

$$\begin{aligned} \widehat{SALES} = {} & 109.72 - 7.640 PRICE + 12.151 ADVERT \\ (\text{se}) \quad & (6.80) \quad (1.046) \qquad (3.556) \\ & -2.768 ADVERT^2 \\ & (0.941) \end{aligned}$$

The t-value for testing $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$ is $t = 7.640/1.045939 = 7.30444$. The 5% critical value for the t-test is $t_c = t_{(0.975,\,71)} = 1.9939$. We reject $H_0: \beta_2 = 0$ because $7.30444 > 1.9939$. The reason for using so many decimals here will soon become clear. We wish to reduce rounding error to ensure the relationship between the t- and F-tests is correctly revealed.

Notice that the squares of the calculated and critical t-values are identical to the corresponding F-values. That is, $t^2 = (7.30444)^2 = 53.355 = F$ and $t_c^2 = (1.9939)^2 = 3.976 = F_c$. The reason for this correspondence is an exact relationship between the t- and F-distributions. The square of a t random variable with df degrees of freedom is an F random variable with 1 degree of freedom in the numerator and df degrees of freedom in the denominator: $t^2_{(df)} = F_{(1,df)}$. Because of this exact relationship, the p-values for the two tests are identical, meaning that we will always reach the same conclusion whichever approach we take. However, there is no equivalence when using a one-tail t-test when the alternative is an inequality such as $>$ or $<$. Because $F = t^2$, the F-test cannot distinguish between the left and right tails as is needed for a one-tail test. Also, the equivalence between t-tests and F-tests does not carry over when a null hypothesis consists of more than a single restriction. Under these circumstances ($J \geq 2$), an F-test needs to be used.

Summarizing the *F*-Test Procedure

1. The null hypothesis H_0 consists of one or more linear equality restrictions on the model parameters β_k. The number of restrictions is denoted by J. When $J = 1$, the null hypothesis is called a single null hypothesis. When $J \geq 2$, it is called a joint null hypothesis. The null hypothesis may not include any "greater than or equal to" or "less than or equal to" hypotheses.
2. The alternative hypothesis states that one or more of the equalities in the null hypothesis is not true. The alternative hypothesis may not include any "greater than" or "less than" options.
3. The test statistic is the F-statistic in equation (6.24).
4. If assumptions MR1–MR6 hold, and if the null hypothesis is true, F has the F-distribution with J numerator degrees of freedom and $N - K$ denominator degrees of freedom. The null hypothesis is *rejected* if $F \geq F_c$, where $F_c = F_{(1-\alpha,\,J,\,N-K)}$ is the critical value that leaves α percent of the probability in the upper tail of the F-distribution.
5. When testing a single equality null hypothesis, it is perfectly correct to use either the t- or F-test procedure: they are equivalent. In practice, it is customary to test single restrictions using a t-test. The F-test is usually reserved for joint hypotheses.

6.1.3 More General *F*-Tests

So far we have discussed the *F*-test in the context of whether a variable or a group of variables could be excluded from the model. The conjectures made in the null hypothesis were that particular coefficients are equal to zero. The *F*-test can also be used for much more general hypotheses. Any number of conjectures ($J \leq K$) involving linear hypotheses with equal signs can be tested. Deriving the restricted model implied by H_0 can be trickier, but the same general principles hold. The restricted sum of squared errors is still greater than the unrestricted sum of squared errors. In the restricted model, least squares estimates are obtained by minimizing the sum of squared errors subject to the restrictions on the parameters being true, and the unconstrained minimum (SSE_U) is always less than the constrained minimum (SSE_R). If SSE_U and SSE_R are substantially different, assuming that the null hypothesis is true significantly reduces the ability of the model to fit the data; in other words, the data do not support the null hypothesis, and it is rejected by the *F*-test. On the other hand, if the null hypothesis is true, we expect the data to be compatible with the conditions placed on the parameters. We expect little change in the sum of squared errors, in which case the null hypothesis will not be rejected by the *F*-test.

EXAMPLE 6.5 | Testing Optimal Advertising

To illustrate how to obtain a restricted model for a null hypothesis that is more complex than assigning zero to a number of coefficients, we return to Example 5.17 where we found that the optimal amount for Andy to spend on advertising $ADVERT_0$ is such that

$$\beta_3 + 2\beta_4 ADVERT_0 = 1 \tag{6.11}$$

Now suppose that Big Andy has been spending \$1900 per month on advertising and he wants to know whether this amount could be optimal. Does the information from the estimated equation provide sufficient evidence to reject a hypothesis that \$1900 per month is optimal? The null and alternative hypotheses for this test are

$$H_0: \beta_3 + 2 \times \beta_4 \times 1.9 = 1 \quad H_1: \beta_3 + 2 \times \beta_4 \times 1.9 \neq 1$$

After carrying out the multiplication, these hypotheses can be written as

$$H_0: \beta_3 + 3.8\beta_4 = 1 \quad H_1: \beta_3 + 3.8\beta_4 \neq 1$$

How do we obtain the restricted model implied by the null hypothesis? Note that when H_0 is true, $\beta_3 = 1 - 3.8\beta_4$. Substituting this restriction into the unrestricted model in (6.9) gives

$$SALES = \beta_1 + \beta_2 PRICE + (1 - 3.8\beta_4)ADVERT + \beta_4 ADVERT^2 + e$$

Collecting terms and rearranging this equation to put it in a form convenient for estimation yields

$$(SALES - ADVERT) = \beta_1 + \beta_2 PRICE + \beta_4 (ADVERT^2 - 3.8ADVERT) + e \tag{6.12}$$

Estimating this model by least squares with dependent variable $y = (SALES - ADVERT)$ and explanatory variables $x_2 = PRICE$ and $x_3 = (ADVERT^2 - 3.8ADVERT)$ yields the restricted sum of squared errors $SSE_R = 1552.286$. The unrestricted sum of squared errors is the same as before, $SSE_U = 1532.084$. We also have one restriction ($J = 1$) and $N - K = 71$ degrees of freedom. Thus, the calculated value of the *F*-statistic is

$$F = \frac{(1552.286 - 1532.084)/1}{1532.084/71} = 0.9362$$

For $\alpha = 0.05$, the critical value is $F_c = 3.976$. Since $F = 0.9362 < F_c = 3.976$, we do not reject H_0. We conclude that Andy's conjecture, that an advertising expenditure of \$1900 per month is optimal is compatible with the data.

Because there is only one conjecture in H_0, you can also carry out this test using the *t*-distribution. Check it out. For the *t*-value, you should find $t = 0.9676$. The value $F = 0.9362$ is equal to $t^2 = (0.9676)^2$, obeying the relationship between *t*- and *F*-random variables that we mentioned previously. You will also find that the *p*-values are identical. Specifically,

$$p\text{-value} = P(F_{(1,71)} > 0.9362) = P(t_{(71)} > 0.9676) + P(t_{(71)} < -0.9676) = 0.3365$$

The result $0.3365 > 0.05$ leads us to conclude that $ADVERT_0 = 1.9$ is compatible with the data.

You may have noticed that our description of this test has deviated slightly from the step-by-step hypothesis testing format introduced in Chapter 3 and used so far in the book.

The same ingredients were there, but the arrangement of them varied. From now on, we will be less formal about following these steps. By being less formal, we can expose you to the type of discussion you will find in research reports, but please remember that the steps were introduced for a purpose: to teach you good habits. Following the steps ensures that you include a description of all the relevant components of the test and that you think about the steps in the correct order. It is **not correct**, for example, to decide on the hypotheses or the rejection region **after** you observe the value of the statistic.

EXAMPLE 6.6 | A One-Tail Test

Suppose that, instead of wanting to test whether the data supports the conjecture "$ADVERT = 1.9$ is optimal," Big Andy wants to test whether the optimal value of $ADVERT$ is greater than 1.9. If he has been spending \$1900 per month on advertising, and he does not want to increase this amount unless there is convincing evidence that the optimal amount is greater than \$1900, he will set up the hypotheses

$$H_0: \beta_3 + 3.8\beta_4 \leq 1 \quad H_1: \beta_3 + 3.8\beta_4 > 1 \tag{6.13}$$

In this case, we can no longer use the F-test. Using a t-test instead, your calculations will reveal $t = 0.9676$. The rejection region for a 5% significane level is reject H_0 if $t \geq 1.667$. Because $0.9676 < 1.667$, we do not reject H_0. There is not enough evidence in the data to suggest the optimal level of advertising expenditure is greater than \$1900.

6.1.4 Using Computer Software

Though it is possible and instructive to compute an F-value by using the restricted and unrestricted sums of squares, it is often more convenient to use the power of econometric software. Most software packages have commands that will automatically compute t- and F-values and their corresponding p-values when provided with a null hypothesis. You should check your software. Can you work out how to get it to test null hypotheses similar to those we constructed? These tests belong to a class of tests called **Wald tests**; your software might refer to them in this way. Can you reproduce the answers we got for all the tests in Chapters 5 and 6?

EXAMPLE 6.7 | Two ($J = 2$) Complex Hypotheses

In this example, we consider a joint test of two of Big Andy's conjectures. In addition to proposing that the optimal level of monthly advertising expenditure is \$1900, Big Andy is planning staffing and purchasing of inputs on the assumption that when $PRICE = \$6$ and $ADVERT = 1.9$, sales revenue will be \$80,000 on average. In the context of our model, and in terms of the regression coefficients β_k, the conjecture is

$$\begin{aligned} E(SALES|PRICE = 6, ADVERT = 1.9) &= \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 \\ &= \beta_1 + 6\beta_2 + 1.9\beta_3 + 1.9^2\beta_4 \\ &= 80 \end{aligned}$$

Are the conjectures about sales and optimal advertising compatible with the evidence contained in the sample of data? We formulate the joint null hypothesis

$$H_0: \beta_3 + 3.8\beta_4 = 1,\ \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 80$$

The alternative is that at least one of these restrictions is not true. Because there are $J = 2$ restrictions to test jointly, we use an F-test. A t-test is not suitable. Note also that this is an example of a test with two restrictions that are more general than simply omitting variables. Constructing the restricted model requires substituting both of these restrictions into our extended model, which is left as an exercise. Using instead computer output obtained by supplying the two hypotheses directly to the software, we obtain a computed value for the F-statistic of 5.74 and a corresponding p-value of 0.0049. At a 5% significance level, the joint null hypothesis is rejected. As another exercise, use the least squares estimates to predict sales revenue for $PRICE = 6$ and $ADVERT = 1.9$. Has Andy been too optimistic about the level of sales, or too pessimistic?

6.1.5 Large Sample Tests

There are two key requirements for the F-statistic to have the F-distribution in samples of all sizes: (1) assumptions MR1–MR6 must hold and (2) the restrictions in H_0 must be *linear* functions of the parameters $\beta_1, \beta_2, \ldots, \beta_K$. In this section, we are concerned with what test statistics are valid in large samples when the errors are no longer normally distributed or when the strict exogeneity assumption is weakend to $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0\ (i \neq j)$. We will also make a few remarks about testing nonlinear hypotheses.

To appreciate the testing alternatives, details about how the F-statistic in (6.4) is constructed are in order. An F random variable is defined as the ratio of two independent chi-square (χ^2) random variables, each divided by their degrees of freedom.[1] That is, if $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$, and V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1, m_2)}$$

In our case, the two independent χ^2 random variables are

$$V_1 = \frac{(SSE_R - SSE_U)}{\sigma^2} \sim \chi^2_{(J)} \quad \text{and} \quad V_2 = \frac{(N-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$$

If σ^2 were known, V_1 would be a natural candidate for testing whether the difference between SSE_R and SSE_U is sufficiently large to reject a null hypothesis. Because σ^2 is unknown, we use V_2 to eliminate it. Specifically,

$$F = \frac{V_1/J}{V_2/(N-K)} = \frac{\dfrac{(SSE_R - SSE_U)}{\sigma^2}\Big/J}{\dfrac{(N-K)\hat{\sigma}^2}{\sigma^2}\Big/(N-K)} = \frac{(SSE_R - SSE_U)/J}{\hat{\sigma}^2} \sim F_{(J, N-K)} \tag{6.13}$$

Note that $\hat{\sigma}^2 = SSE_U/(N-K)$, and so the result in (6.13) is identical to the F-statistic first introduced in (6.4).

When we drop the normality assumption or weaken the strict exogeneity assumption, the argument becomes slightly different. In this case, V_1 no longer has an exact χ^2-distribution, but we can nevertheless rely on asymptotic theory to say that

$$V_1 = \frac{(SSE_R - SSE_U)}{\sigma^2} \overset{a}{\sim} \chi^2_{(J)}$$

Then, we can go one step further and say that replacing σ^2 by its consistent estimator $\hat{\sigma}^2$ does not change the asymptotic distribution of V_1.[2] That is,

$$\hat{V}_1 = \frac{(SSE_R - SSE_U)}{\hat{\sigma}^2} \overset{a}{\sim} \chi^2_{(J)} \tag{6.14}$$

This statistic is a valid alternative for testing joint linear hypotheses in large samples under less restrictive assumptions, with the approximation improving as sample size increases. At a 5% significance level, we reject H_0 if $\hat{V}_1$ is greater than or equal to the critical value $\chi^2_{(0.95,J)}$, or if the p-value $P\left(\chi^2_{(J)} > \hat{V}_1\right)$ is less than 0.05. In response to an automatic test command, most software will give you values for both F and $\hat{V}_1$. The value for $\hat{V}_1$ will probably be referred to as "chi-square."

Although it is clear that $F = \hat{V}_1/J$, the two test alternatives will not necessarily lead to the same outcome; their p-values will be different. Both are used in practice, and it is possible

[1] See Appendices B.3.6 and B.3.8.

[2] See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorem D.16, page 1168 of online Appendix.

that the F-test will provide a better small-sample approximation than $\hat{V}_1$ even under the less restrictive assumptions. As the sample size grows (the degrees of freedom for the denominator of the F-statistic increase), the two tests become identical—their p-values become the same, and their critical values beome equivalent in the sense that $\lim_{N\to\infty} F_{(1-\alpha,J,N-K)} = \chi^2_{(1-\alpha,J)}/J$. Check it out yourself. Suppose $J = 4$ and $\alpha = 0.05$, then from Statistical Table 3, $\chi^2_{(0.95,4)}/4 = 9.488/4 = 2.372$. The F-values are in Statistical Table 4, but it is instructive to use software to provide a few extra values. Doing so, we find $F_{(0.95,\,4,\,60)} = 2.525$, $F_{(0.95,\,4,\,120)} = 2.447$, $F_{(0.95,\,4,\,500)} = 2.390$, $F_{(0.95,\,4,\,1000)} = 2.381$, and $F_{(0.95,\,4,\,10000)} = 2.373$. As $N - K$ increases, the 95th percentile of the F-distribution approaches 2.372.

EXAMPLES 6.2 and 6.5 | Revisited

When testing $H_0: \beta_3 = \beta_4 = 0$ in the equation

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.15}$$

we obtain $F = 8.44$ with corresponding p-value $= 0.0005$, and $\chi^2 = 16.88$ with corresponding p-value $= 0.0002$. Because there are two restrictions ($J = 2$), the F-value is half the χ^2-value. The p-values are different because the tests are different.

For testing $H_0: \beta_3 + 3.8\beta_4 = 1$, we obtain $F = 0.936$ with corresponding p-value $= 0.3365$ and $\chi^2 = 0.936$ with corresponding p-value $= 0.3333$. The F- and χ^2-values are equal because $J = 1$, but again the p-values are slightly different.

Testing Nonlinear Hypotheses

Test statistics for joint hypotheses which are nonlinear functions of the parameters are more challenging theoretically,[3] but nevertheless can typically be carried out by your software with relative ease. Only asymptotic results are available, and the relevant test statistic is the chi-square, although you may find that some software also gives an F-value. Another thing to be on lookout for is whether a nonlinear hypothesis can be re-framed as a linear hypothesis to avoid one aspect of the approximation.

EXAMPLE 6.8 | A Nonlinear Hypothesis

In Section 5.7.4, we found that, in terms of the parameters of equation (6.2), the optimal level of advertising is given by

$$ADVERT_0 = \frac{1-\beta_3}{2\beta_4}$$

To test the hypothesis that the optimal level is \$1,900 against the alternative that it is not \$1,900, we can set up the following hypotheses which are nonlinear in the parameters

$$H_0: \frac{1-\beta_3}{2\beta_4} = 1.9 \quad H_1: \frac{1-\beta_3}{2\beta_4} \neq 1.9 \tag{6.16}$$

There are three ways we can approach this problem. The first way is to convert the hypotheses so that they are linear in the parameters. That is, $H_0: \beta_3 + 3.8\beta_4 = 1$ versus $H_1: \beta_3 + 3.8\beta_4 \neq 1$. These are the hypotheses that we tested in Example 6.5. The p-value for the F-test was 0.337.

The second way is to test (6.16) using the t-test value

$$t = \frac{g(b_3, b_4) - 1.9}{\text{se}\left[g(b_3, b_4)\right]}$$

$$= \frac{(1-b_3)/2b_4 - 1.9}{\text{se}\left((1-b_3)/2b_4\right)} = \frac{2.0143 - 1.9}{0.1287} = 0.888$$

The values $g(b_3, b_4) = (1-b_3)/2b_4 = 2.0143$ and $\text{se}\left[g(b_3, b_4)\right] = \text{se}\left((1-b_3)/2b_4\right) = 0.1287$, were found in Example 5.17 for computing an interval estimate for $ADVERT_0$. The third way is to use the χ^2-test for testing (6.16). When we have only a single hypothesis, $\chi^2 = F = t^2 = (0.888)^2 = 0.789$. The F and t^2 critical values correspond, yielding a p-value of 0.377. The χ^2-test is a different test, however. It yields a p-value of 0.374.

[3] See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, pp. 211–212.

Having so many options will undoubtedly leave you wondering what to do. In general, the best strategy is to convert the hypotheses into ones that are linear if that is possible. Otherwise, the t- or χ^2-tests can be used, but the t-test option is not available if $J \geq 2$. The important thing to take away from this section is an appreciation of the different test statistics that appear on your software output—what they mean, where they come from, and the circumstances under which they are exact finite sample tests or asymptotic approximations.

6.2 The Use of Nonsample Information

In many estimation problems we have information over and above the information contained in the sample observations. This nonsample information may come from many places, such as economic principles or experience. When it is available, it seems intuitive that we should find a way to use it. If the nonsample information is correct, and if we combine it with the sample information, the precision with which we can estimate the parameters is improved.

To illustrate how we might go about combining sample and nonsample information, consider a model designed to explain the demand for beer. From the theory of consumer choice in microeconomics, we know that the demand for a good will depend on the price of that good, on the prices of other goods—particularly substitutes and complements—and on income. In the case of beer, it is reasonable to relate the quantity demanded (Q) to the price of beer (PB), the price of liquor (PL), the price of all other remaining goods and services (PR), and income (I). To estimate this demand relationship, we need a further assumption about the functional form. Using "ln" to denote the natural logarithm, we assume, for this case, that the log-log functional form is appropriate:

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + e \tag{6.17}$$

This model is a convenient one because it precludes infeasible negative prices, quantities, and income, and because the coefficients β_2, β_3, β_4, and β_5 are elasticities. See Section 4.6.

A relevant piece of nonsample information can be derived by noting that if all prices and income go up by the same proportion, we would expect there to be no change in quantity demanded. For example, a doubling of all prices and income should not change the quantity of beer consumed. This assumption is that economic agents do not suffer from "money illusion." Let us impose this assumption on our demand model and see what happens. Having all prices and income change by the same proportion is equivalent to multiplying each price and income by a constant. Denoting this constant by λ and multiplying each of the variables in (6.17) by λ yields

$$\begin{aligned} \ln(Q) &= \beta_1 + \beta_2 \ln(\lambda PB) + \beta_3 \ln(\lambda PL) + \beta_4 \ln(\lambda PR) + \beta_5 \ln(\lambda I) \\ &= \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) \\ &\quad + \left(\beta_2 + \beta_3 + \beta_4 + \beta_5\right) \ln(\lambda) + e \end{aligned} \tag{6.18}$$

Comparing (6.17) with (6.18) shows that multiplying each price and income by λ will give a change in $\ln(Q)$ equal to $\left(\beta_2 + \beta_3 + \beta_4 + \beta_5\right)\ln(\lambda)$. Thus, for there to be no change in $\ln(Q)$ when all prices and income go up by the same proportion, it must be true that

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0 \tag{6.19}$$

Thus, we can say something about how quantity demanded should not change when prices and income change by the same proportion, and this information can be written in terms of a specific restriction on the parameters of the demand model. We call such a restriction **nonsample information**. If we believe that this nonsample information makes sense, and hence that the parameter restriction in (6.19) holds, then it seems desirable to be able to obtain estimates that obey this restriction.

To introduce the nonsample information, we solve the parameter restriction $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$ for one of the β_k's. Which one is not important mathematically, but for reasons that will become apparent, we solve for β_4:

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5$$

Substituting this expression into the original model in (6.17) gives

$$\begin{aligned}\ln(Q) &= \beta_1 + \beta_2\ln(PB) + \beta_3\ln(PL) + \left(-\beta_2 - \beta_3 - \beta_5\right)\ln(PR) + \beta_5\ln(I) + e \\ &= \beta_1 + \beta_2\left[\ln(PB) - \ln(PR)\right] + \beta_3\left[\ln(PL) - \ln(PR)\right] + \beta_5\left[\ln(I) - \ln(PR)\right] + e \\ &= \beta_1 + \beta_2\ln\left(\frac{PB}{PR}\right) + \beta_3\ln\left(\frac{PL}{PR}\right) + \beta_5\ln\left(\frac{I}{PR}\right) + e \end{aligned} \tag{6.20}$$

By using the restriction to replace β_4, and using the properties of logarithms, we have constructed the new variables $\ln(PB/PR)$, $\ln(PL/PR)$, and $\ln(I/PR)$. These variables have an appealing interpretation. Because PR represents the price of all other goods and services, (PB/PR) and (PL/PR) can be viewed as the *real* price of beer and the *real* price of liquor, respectively, and (I/PR) can be viewed as *real* income. By applying least squares to the restricted equation (6.20), we obtain the **restricted least squares estimates** $\left(b_1^*, b_2^*, b_3^*, b_5^*\right)$. The restricted least squares estimate for β_4 is given by $b_4^* = -b_2^* - b_3^* - b_5^*$.

EXAMPLE 6.9 | Restricted Least Squares

Observations on Q, PB, PL, PR, and I, taken from a cross section of 30 households are stored in the file *beer*. Using these observations to estimate (6.20), we obtain

$$\begin{aligned}\widehat{\ln(Q)} = &-4.798 - \underset{(0.166)}{1.2994}\ln\left(\frac{PB}{PR}\right) + \underset{(0.284)}{0.1868}\ln\left(\frac{PL}{PR}\right) \\ &+ \underset{(0.427)}{0.9458}\ln\left(\frac{I}{PR}\right)\end{aligned}$$

(se)

and $b_4^* = -(-1.2994) - 0.1868 - 0.9458 = 0.1668$. We estimate the price elasticity of demand for beer as -1.30, the cross-price elasticity of demand for beer with respect to liquor as 0.19, the cross-price elasticity of demand for beer with respect to other goods and services as 0.17, and the income elasticity of demand for beer as 0.95.

Substituting the restriction into the original equation and rearranging it like we did to get (6.20) will always work, but it may not be necessary. Different software has different options for obtaining restricted least squares estimates. Please check what is available in the software of your choice.

What are the properties of the restricted least squares estimation procedure? If assumptions MR1–MR5 hold for the unrestricted model, then the restricted least squares estimator is biased, $E\left(b_k^*\right) \neq \beta_k$, *unless* the constraints we impose are *exactly* true. This result makes an important point about econometrics. A good *economist* will obtain more reliable parameter estimates than a poor one because a good economist will introduce better nonsample information. This is true at the time of model specification as well as later, when constraints might be applied to the model. Nonsample information is not restricted to constraints on the parameters; it is also used for model specification. *Good economic theory* is a very important ingredient in empirical research.

The second property of the restricted least squares estimator is that its variance is smaller than the variance of the least squares estimator, *whether the constraints imposed are true or not*. By combining nonsample information with the sample information, we reduce the variation in the estimation procedure caused by random sampling. This reduction in variance obtained by imposing restrictions on the parameters is not at odds with the Gauss–Markov theorem. The Gauss–Markov result that the least squares estimator is the best linear unbiased estimator applies

to linear and unbiased estimators that use data alone, and no constraints on the parameters. Including additional information with the data gives the added reward of a reduced variance. If the additional nonsample information is correct, we are unambiguously better off; the restricted least squares estimator is unbiased and has lower variance. If the additional nonsample information is incorrect, the reduced variance comes at the cost of bias. This bias can be a big price to pay if it leads to estimates substantially different from their corresponding true parameter values. Evidence on whether or not a restriction is true can be obtained by testing the restriction along the lines of the previous section. In the case of this particular demand example, the test is left as an exercise.

6.3 Model Specification

In what has been covered so far, we have generally taken the role of the model as given. Questions have been of the following type: Given a particular regression model, what is the best way to estimate its parameters? Given a particular model, how do we test hypotheses about the parameters of that model? How do we construct interval estimates for the parameters of a model? What are the properties of estimators in a given model? Given that all these questions require knowledge of the model, it is natural to ask where the model comes from. In any econometric investigation, choice of the model is one of the first steps. In this section, we focus on the following questions: What are the important considerations when choosing a model? What are the consequences of choosing the wrong model? Are there ways of assessing whether a model is adequate?

Three essential features of model choice are (1) choice of functional form, (2) choice of explanatory variables (regressors) to be included in the model, and (3) whether the multiple regression assumptions MR1–MR6, listed in Chapter 5, hold. The implications of some violations of these assumptions have already been discussed. In particular, we have seen how it is necessary to rely on large sample results for inference if the errors are no longer normally distributed (MR6 is violated), or if assumption MR2: $E(e_i|\mathbf{X}) = 0$ is weakened to the alternative assumption that $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0$ for $i \neq j$. Later chapters on heteroskedasticity, regression with time-series data, and endogenous regressors deal with violations of MR3, MR4 and $\text{cov}(e_i, x_{jk}) = 0$. In this section, we focus mainly on issues dealing with choice of regressors and also give some consideration to choice of functional form. The properties of alternative functional forms were considered in Sections 2.8, 4.3–4.6, and 5.6. When making a functional-form choice, we need to ask questions such as: How is the dependent variable y likely to respond when the regressors change? At a constant rate? At a decreasing rate? Is it reasonable to assume constant elasticities over the whole range of the data? Are there any patterns in the least squares residuals that suggest an alternative functional form? The use of least squares residuals for assessing the adequacy of a functional form was considered in Section 4.3.4.

For choice of regressors, a fundamental consideration is the purpose of the model—whether it is intended for prediction or for causal analysis. We turn now to that question.

6.3.1 Causality versus Prediction

With causal inference we are primarily interested in the effect of a change in a regressor on the conditional mean of the dependent variable. Is there an effect and, if so, what is its magnitude? We wish to be able to say that a one-unit change in an explanatory variable will cause a particular change in the mean of the dependent variable, other factors held constant. This type of analysis is important for policy work. For example, suppose a government is concerned about educational performance in schools and believes that large class sizes may be the cause of poor performance. Before it spends large sums of money increasing the number of teachers, and building more classrooms, it would want convincing evidence that class size does have an impact on performance. We would need to be able to separate the effect of class size from the effect of other variables

such as socioeconomic background. It may be that large classes tend to be in areas of poor socioeconomic background. Under these circumstances it is important to include all relevant variables so that we can be sure "other factors are held constant" when we measure the effect of class size.

On the other hand, if the purpose of a model is to predict the value of a dependent variable, then, for regressor choice, it is important to choose variables that are highly correlated with the dependent variable and that lead to a high R^2. Whether or not these variables have a direct effect on the dependent variable, and the possible omission of some relevant variables, are less important. Predictive analysis using variables from the increasingly popular field of "big data" is an example of where variables are chosen for their predictive ability rather than to examine causal relationships.

To appreciate the difference in emphasis, and when it matters, suppose the variables $(y_i, x_i, z_i), i = 1, 2, \ldots, N$ are randomly selected from a population satisfying

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i \tag{6.21}$$

We have chosen the notation x for one of the explanatory variables and z for the other explanatory variable to distinguish between what will be an included variable x and an omitted variable z. We assume $E(e_i|x_i, z_i) = 0$ and hence $E(y_i|x_i, z_i) = \beta_1 + \beta_2 x_i + \beta_3 z_i$. Under these assumptions, β_2 and β_3 have the causal interpretations

$$\beta_2 = \frac{\partial E(y_i|x_i, z_i)}{\partial x_i} \quad \beta_3 = \frac{\partial E(y_i|x_i, z_i)}{\partial z_i}$$

That is, β_2 represents the change in the mean of y from a change in x, other factors held constant, and β_3 represents the change in the mean of y from a change in z, other factors held constant The assumption $E(e_i|x_i, z_i) = 0$ is important for these interpretations. It means that changes in x_i or z_i have no impact on the error term. Now suppose that x_i and z_i are correlated, a common occurrence among explanatory variables. Because they are correlated, $E(z_i|x_i)$ will depend on x_i. Let us assume that this dependence can be represented by the linear function

$$E(z_i|x_i) = \gamma_1 + \gamma_2 x_i \tag{6.22}$$

Then, using (6.21) and (6.22), we have

$$\begin{aligned} E(y_i|x_i) &= \beta_1 + \beta_2 x_i + \beta_3 E(z_i|x_i) + E(e_i|x_i) \\ &= \beta_1 + \beta_2 x_i + \beta_3(\gamma_1 + \gamma_2 x_i) \\ &= (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2)x_i \end{aligned}$$

where $E(e_i|x_i) = E_z[E(e_i|x_i, z_i)] = 0$ by the law of iterated expectations. If knowing x_i or z_i does not help to predict e_i, then knowing x_i does not help to predict e_i either.

Now, we can define $u_i = y_i - E(y_i|x_i)$, $\alpha_1 = \beta_1 + \beta_3\gamma_1$, and $\alpha_2 = \beta_2 + \beta_3\gamma_2$, and write

$$\begin{aligned} y_i &= (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2)x_i + u_i \\ &= \alpha_1 + \alpha_2 x_i + u_i \end{aligned} \tag{6.23}$$

where $E(u_i|x_i) = 0$ by definition. Application of least squares to (6.23) will yield best linear unbiased estimates of α_1 and α_2. If the objective is to use x_i to predict y_i, we can proceed with this equation without worrying about the omission of z_i. However, because z_i is not held constant, α_2 does not measure the causal effect of x_i on y_i, which is given by β_2. The coefficient α_2 includes the indirect effect of x_i on z_i through γ_2 (which may or may not be causal), followed by the effect of that change in z_i on y_i, through β_3. Note that if $\beta_3 = 0$ (z_i does not effect y_i) or $\gamma_2 = 0$ (z_i and x_i are uncorrelated), then $\alpha_2 = \beta_2$ and estimation of α_2 gives the required causal effect.

Thus, to estimate a causal effect of a variable x using least squares, we need to start with a model where all variables that are correlated with x and impact on y are included. An alternative, valuable when data on all such variables are not available, is to use control variables. We discuss their use in Section 6.3.4.

6.3.2 Omitted Variables

As explained in the previous section, if our objective is to estimate a causal relationship, then the possible omission of relevant variables is a concern. In this section, we explore further the impact of omitting important variables. Such omissions are always a possibility. Our economic principles may have overlooked a variable, or lack of data may lead us to drop a variable even when it is prescribed by economic theory.

EXAMPLE 6.10 | Family Income Equation

To introduce the **omitted variable problem**, we consider a sample of married couples where both husbands and wives work. This sample was used by labor economist Tom Mroz in a classic paper on female labor force participation. The variables from this sample that we use in our illustration are stored in the file *edu_inc*. The dependent variable is the logarithm of annual family income *FAMINC* defined as the combined income of husband and wife. We are interested in the impact of level of education, both the husband's education (*HEDU*) and the wife's education (*WEDU*), on family income. The first equation to be estimated is

$$\ln(FAMINC) = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + e \qquad (6.24)$$

Coefficient estimates from this equation, their standard errors, and their *p*-values for testing whether they are significantly different from zero, are given in column (1) of Table 6.1. We estimate that an additional year of education

TABLE 6.1 Estimated Equations for Family Income

	ln(*FAMINC*)				
	(1)	(2)	(3)	(4)	(5)
C	10.264	10.539	10.238	10.239	10.310
HEDU	0.0439	0.0613	0.0448	0.0460	0.0517
(se)	(0.0087)	(0.0071)	(0.0086)	(0.0136)	(0.0133)
[*p*-value]	[0.0000]	[0.0000]	[0.0000]	[0.0007]	[0.0001]
WEDU	0.0390		0.0421	0.0492	
(se)	(0.0116)		(0.0115)	(0.0247)	
[*p*-value]	[0.0003]		[0.0003]	[0.0469]	
KL6			–0.1733	–0.1724	–0.1690
(se)			(0.0542)	(0.0547)	(0.0548)
[*p*-value]			[0.0015]	[0.0017]	[0.0022]
XTRA_X5				0.0054	–0.0321
(se)				(0.0243)	(0.0154)
[*p*-value]				[0.8247]	[0.0379]
XTRA_X6				–0.0069	0.0309
(se)				(0.0215)	(0.0101)
[*p*-value]				[0.7469]	[0.0023]
SSE	82.2648	84.4623	80.3297	80.3062	81.0622
RESET *p*-values					
1 term $(\hat{y}^2)$	0.3374	0.1017	0.1881	0.1871	0.1391
2 terms $(\hat{y}^2, \hat{y}^3)$	0.1491	0.0431	0.2796	0.2711	0.2715

for the husband will increase annual income by 4.4%, and an additional year of education for the wife will increase income by 3.9%. Both estimates are significantly different from zero at a 1% level of significance.[4]

What happens if we now incorrectly omit wife's education from the equation? The resulting estimates are given in column (2) of Table 6.1. Omitting *WEDU* leads to an estimate that suggests the effect of an extra year of education for the husband is 6.1%. The effect of the wife's education has been incorrectly attributed to the husband's education leading to an overstatement of the latter's importance. This change in the magnitude of a coefficient is typical of the effect of incorrectly omitting a relevant variable. Omission of a relevant variable (defined as one whose coefficient is nonzero) leads to an estimator that is biased. Naturally enough, this bias is known as **omitted variable bias**.

Omitted Variable Bias: A Proof To give a general expression for the bias for the case where one explanatory variable is omitted from a model with two explanatory variables, consider the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$. Suppose that we incorrectly omit z_i from the model and estimate instead $y_i = \beta_1 + \beta_2 x_i + v_i$ where $v_i = \beta_3 z_i + e_i$. Then, the estimator used for β_2 is

$$b_2^* = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \beta_2 + \sum w_i v_i$$

where $w_i = (x_i - \bar{x}) / \sum(x_i - \bar{x})^2$. The second equality in this equation follows from Appendix 2D. Substituting for v_i yields

$$b_2^* = \beta_2 + \beta_3 \sum w_i z_i + \sum w_i e_i$$

Assuming that $E(e_i|\mathbf{x}, \mathbf{z}) = 0$, or alternatively, that (y_i, x_i, z_i) is a random sample and $E(e_i|x_i, z_i) = 0$, the conditional mean of b_2^* is

$$E(b_2^*|\mathbf{x}, \mathbf{z}) = \beta_2 + \beta_3 \sum w_i z_i = \beta_2 + \beta_3 \frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)} \tag{6.25}$$

You are asked to prove this result in Exercise 6.3. Unconditionally, we have

$$E(b_2^*) = \beta_2 + \beta_3 E\left[\frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)}\right] \tag{6.26}$$

and in large samples, under less restrictive conditions,

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{cov}(x, z)}{\text{var}(x)} \tag{6.27}$$

Thus, $E(b_2^*) \neq \beta_2$ and b_2^* is not a consistent estimator for β_2. It is biased in small and large samples if $\text{cov}(x, z) \neq 0$. In terms of (6.25)—the result is similar for (6.26) and (6.27)—the bias is given by

$$\text{bias}(b_2^*|\mathbf{x}, \mathbf{z}) = E(b_2^*|\mathbf{x}, \mathbf{z}) - \beta_2 = \beta_3 \frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)} \tag{6.28}$$

We can make four more interesting observations from the results in (6.25)–(6.28).

1. Omitting a relevant variable is a special case of using a restricted least squares estimator where the restriction $\beta_3 = 0$ is not true. It leads to a biased estimator for β_2 but one with a lower variance. In columns (1) and (2) of Table 6.1 the reduction in the standard error for the coefficient of *HEDU* from 0.0087 to 0.0071 is consistent with the lower-variance result.
2. Knowing the sign of β_3 and the sign of the covariance between x and z tells us the direction of the bias. In Example 6.9 we expect a wife's level of education to have a positive effect on family income $(\beta_3 > 0)$, and we expect husband's and wife's levels of education to be

[4]There are a number of other entries in Table 6.1 that we discuss in due course: estimates from other equations and RESET values.

TABLE 6.2 **Correlation Matrix for Variables Used in Family Income Example**

	ln(*FAMINC*)	*HEDU*	*WEDU*	*KL6*	*XTRA_X5*	*XTRA_X6*
ln(*FAMINC*)	1.000					
HEDU	0.386	1.000				
WEDU	0.349	0.594	1.000			
KL6	–0.085	0.105	0.129	1.000		
XTRA_X5	0.315	0.836	0.518	0.149	1.000	
XTRA_X6	0.364	0.821	0.799	0.160	0.900	1.000

positively correlated ($\text{cov}(x, z) > 0$). Thus, we expect an upward bias for the coefficient estimate in (2), as indeed has occurred. The positive correlation between *HEDU* and *WEDU* can be confirmed from the correlation matrix in Table 6.2.

3. The bias in (6.28) can also be written as $\beta_3\hat{\gamma}_2$ where $\hat{\gamma}_2$ is the least squares estimate of γ_2 from the regression equation $E(z|x) = \gamma_1 + \gamma_2 x$. This result is consistent with equation (6.23) where we explained how omitting a relevant variable can lead to an incorrect estimate of a causal effect.
4. The importance of the assumption $E(e_i|\mathbf{x}, \mathbf{z}) = 0$ becomes clear. In the equation $y_i = \beta_1 + \beta_2 x_i + v_i$, we have $E(v_i|x_i) = \beta_3 E(z_i|x_i)$. It is the nonzero value for $E(z_i|x_i)$ that leads to the biased estimator for β_2.

EXAMPLE 6.11 | Adding Children Aged Less Than 6 Years

There are, of course, other variables that could be included as explanators of family income. In column (3) of Table 6.1 we include *KL6*, the number of children less than 6 years old. The larger the number of young children, the fewer the number of hours likely to be worked and hence a lower family income would be expected. The estimated coefficient on *KL6* is negative, confirming this expectation. Also, despite the fact that *KL6* is not highly correlated with *HEDU* and *WEDU*, the coefficient estimates for these variables have increased slightly, indicating that once we hold the number of young children constant, the returns to education for both the wife and the husband are greater, with the greater increase going to the wife whose working hours would be the more likely to be affected by the presence of young children.

6.3.3 Irrelevant Variables

The consequences of omitting relevant variables may lead you to think that a good strategy is to include as many variables as possible in your model. However, doing so will not only complicate your model unnecessarily, it may inflate the variances of your estimates because of the presence of **irrelevant variables**—those whose coefficients are zero because they have no direct effect on the dependent variable.

EXAMPLE 6.12 | Adding Irrelevant Variables

To see the effect of irrelevant variables, we add two artificially generated variables *XTRA_X5* and *XTRA_X6* to the family income equation. These variables were constructed so that they are correlated with *HEDU* and *WEDU* but have no influence on family income. The results from including these two variables are given in column (4) of Table 6.1. What can we observe from these estimates? First, as expected, the coefficients of *XTRA_X5* and *XTRA_X6*

have p-values greater than 0.05. They do indeed appear to be irrelevant variables. Also, the standard errors of the coefficients estimated for all other variables have increased, with p-values increasing correspondingly. The inclusion of irrelevant variables that are correlated with the other variables in the equation has reduced the precision of the estimated coefficients of the other variables. This result follows because, by the Gauss–Markov theorem, the least squares estimator of the correct model is the minimum variance linear unbiased estimator.

Finally, let us check what happens if we retain *XTRA_X5* and *XTRA_X6*, but omit *WEDU*, leading to the results in column (5). The coefficients for *XTRA_X5* and *XTRA_X6* have become significantly different from zero at a 5% level of significance. The irrelevant variables have picked up the effect of the relevant omitted variable. While this may not matter if prediction is the main objective of the exercise, it can lead to very erroneous conclusions if we are trying to identify the causal effects of the included variables.

6.3.4 Control Variables

In the discussion so far, we have not explicitly distinguished between variables whose causal effect is of particular interest and other variables that may simply be in the equation to avoid omitted variable bias in the estimate of the causal coefficient. Variables included in the equation to avoid omitted variable bias in the coefficient of interest are called **control variables**. Control variables may be included in the equation because they have a direct effect on the dependent variable in their own right or because they can act as proxy variables for relevant omitted variables that are difficult to observe. For a control variable to serve its purpose and act as a proxy for an omitted variable, it needs to satisfy a **conditional mean independence** assumption. To introduce this assumption, we return to the equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i \tag{6.29}$$

where the observation (y_i, x_i, z_i) is obtained by random sampling and where $E(e_i|x_i, z_i) = 0$. Suppose we are interested in β_2, the causal effect of x_i on y_i, and, although β_3 provides the causal effect of z_i on y_i, we are not concerned about estimating it. Also suppose that z_i is omitted from the equation because it is unobservable or because data on it are too difficult to obtain, leaving the equation

$$y_i = \beta_1 + \beta_2 x_i + v_i$$

where $v_i = \beta_3 z_i + e_i$. If z_i and x_i are uncorrelated, there are no problems. Application of least squares to $y_i = \beta_1 + \beta_2 x_i + v_i$ will yield a consistent estimate for β_2. However, as indicated in (6.28), correlation between z_i and x_i leads to a bias in the least squares estimator for β_2 equal to $\beta_3 \text{cov}(x, z)/\text{var}(x)$.

Now consider another variable q that has the property

$$E(z_i|x_i, q_i) = E(z_i|q_i) \tag{6.30}$$

This property says that once we know q, knowing x does not provide any more information about z. It means that x and z will no longer be correlated once q has been partialled out. We say that z_i and x_i are **conditionally mean independent**. An example will help cement this concept.

When labor economists estimate wage equations they are particularly interested in the returns to education. In particular, what is the causal relationship between more education and higher wages? Other variables such as experience are typically added to the equation, but they are usually not the main focus. One variable that is clearly relevant, but difficult to include because it cannot be observed, is ability. Also, more able people are likely to have more education, and so ability and education will be correlated. Excluding the variable "ability" will bias the estimate of the causal effect of education on wages. Suppose, however, that we have observations on *IQ*. *IQ* will clearly be correlated with both education and ability. Will it satisfy the conditional mean independence assumption? We need to be able to write

$$E(ABILITY|EDUCATION, IQ) = E(ABILITY|IQ)$$

That is, once we know somebody's *IQ*, knowing their level of education does not add any extra information about their ability. Another way to think about it is that education is "as if" it was randomly assigned, once we have taken *IQ* into account. One could argue whether this is a reasonable assumption, but, if it is reasonable, then we can proceed to use *IQ* as a control variable or a proxy variable to replace *ABILITY*.

How a Control Variable Works Returning to equation (6.29), namely, $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, we can write

$$E(y_i|x_i, q_i) = \beta_1 + \beta_2 x_i + \beta_3 E(z_i|x_i, q_i) + E(e_i|x_i, q_i) \tag{6.31}$$

If the conditional mean independence assumption in (6.30) holds, then $E(z_i|x_i, q_i) = E(z_i|q_i)$. For illustrative purposes, we assume $E(z_i|q_i)$ is a linear function of q_i, say $E(z_i|q_i) = \delta_1 + \delta_2 q_i$. We also need to assume that q_i has no direct effect on y_i, so that $E(e_i|x_i, q_i) = 0$.[5] Inserting these results into (6.31), we have

$$\begin{aligned} E(y_i|x_i, q_i) &= \beta_1 + \beta_2 x_i + \beta_3(\delta_1 + \delta_2 q_i) \\ &= \beta_1 + \beta_3\delta_1 + \beta_2 x_i + \beta_3\delta_2 q_i \\ &= \alpha_1 + \beta_2 x_i + \alpha_2 q_i \end{aligned}$$

where $\alpha_1 = \beta_1 + \beta_3\delta_1$ and $\alpha_2 = \beta_3\delta_2$. Defining $u_i = y_i - E(y_i|x_i, q_i)$, we have the equation

$$y_i = \alpha_1 + \beta_2 x_i + \alpha_2 q_i + u_i$$

Since $E(u_i|x_i, q_i) = 0$ by definition, least squares estimates of α_1, β_2, and α_2 will be consistent. Notice that we have been able to estimate β_2, the causal effect of x on y, but we have not been able to consistently estimate β_3, the causal effect of z on y.

This result holds if q is a perfect proxy for z. We may want to ask what happens if the conditional mean independence assumption does not hold, making q an **imperfect proxy** for z. Suppose

$$E(z_i|x_i, q_i) = \delta_1 + \delta_2 q_i + \delta_3 x_i$$

In this case q is not a perfect proxy because, after controlling for it, $E(z_i|x_i, q_i)$ still depends on x. Using similar algebra, we obtain

$$E(y_i|x_i, q_i) = (\beta_1 + \beta_3\delta_1) + (\beta_2 + \beta_3\delta_3) x_i + \beta_3\delta_2 q_i$$

The bias from using this equation to estimate β_2 is $\beta_3\delta_3$. The bias from omitting z instead of using the control variable is $\beta_3 \text{cov}(x, z)/\text{var}(x)$. Thus, for the control variable to be an improvement over omission of z, we require $\delta_3 < \text{cov}(x, z)/\text{var}(x)$. Now, $\text{cov}(x, z)/\text{var}(x)$ is equal to the coefficient of x in a regression of z on x. Thus, the condition $\delta_3 < \text{cov}(x, z)/\text{var}(x)$ is equivalent to saying that the coefficient of x in a regression of z on x is lower after the inclusion of q. Put another way, after partialling out q, the correlation between x and z is reduced but not eliminated.

EXAMPLE 6.13 | A Control Variable for Ability

To illustrate the use of a control variable, we consider the model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 ABILITY + e$$

and use data stored in the data file *koop_tobias_87*, a subset of data used by Koop and Tobias.[6] The sample is restricted to white males who are at least 16 years of age and who worked at least 30 weeks and 800 hours during the year. The Koop–Tobias data extend from 1979 to 1993. We use

[5] In Exercise 6.4 you are invited to investigate how this assumption can be relaxed.

[6] G. Koop and J.L. Tobias (2004), "Learning about Heterogeneity in Returns to Schooling", *Journal of Applied Econometrics*, 19, 827–849.

observations from 1987, a total of $N = 1057$. The variables *EDUC* and *EXPER* are numbers of years of education and experience, respectively. The variable *ABILITY* is unobserved, but we have instead the proxy variable *SCORE*, which is constructed from the 10 component tests of the Armed Services Vocational Aptitude Battery, administered in 1980, and standardized for age. Omitting *ABILITY*, the least squares estimated equation is

$$\begin{aligned}\widehat{\ln(WAGE)} &= 0.887 + 0.0728EDUC + 0.01268EXPER \\ \text{(se)} \quad & (0.293)\ (0.0091) \qquad (0.0403) \\ & - 0.00571EXPER^2 \\ & (0.00165)\end{aligned}$$

Including the proxy variable *SCORE*, we obtain

$$\begin{aligned}\widehat{\ln(WAGE)} &= 1.055 + 0.0592EDUC + 0.1231EXPER \\ \text{(se)} \quad & (0.297)\ (0.0101) \qquad (0.0401) \\ & - 0.00538EXPER^2 + 0.0604SCORE \\ & (0.00165) \qquad (0.0195)\end{aligned}$$

The return to an extra year of education drops from 7.3% to 5.9% after including the variable *SCORE*, suggesting that omitting the variable *ABILITY* has incorrectly attributed some of its effect to the level of education. There has been little effect on the coefficients of *EXPER* and $EXPER^2$. The conditional mean independence assumption that has to hold to conclude that extra *EDUC* causes a 5.9% increase in *WAGE* is $E(ABILITY|EDUC, EXPER, SCORE) = E(ABILITY|EXPER, SCORE)$. After allowing for *EXPER* and *SCORE*, knowing *EDUC* does not provide any more information about *ABILITY*. This assumption is required for both the education and experience coefficients to be given a causal interpretation. Finally, we note that the coefficient of the proxy variable *SCORE* cannot be given a causal interpretation.

6.3.5 Choosing a Model

Although choosing a model is fundamental, it is often not an easy task. There is no one set of mechanical rules that can be applied to come up with the best model. The choice will depend on the purpose of the model and how the data were collected, and requires an intelligent application of both theoretical knowledge and the outcomes of various statistical tests. Better choices come with experience. What is important is to recognize ways of assessing whether a model is reasonable or not. The following points are helpful for such an assessment.

1. Is the purpose of the model to identify one or more causal effects or is it prediction? Where causality is the focus, omitted variable bias can invalidate conclusions. Careful selection of control variables, whether they be variables in their own right or proxy variables, is necessary. On the other hand, if prediction is the objective, then the major concern is using variables that have high predictive power because of their correlation with the dependent variable. Omitted variables bias is not a major concern.
2. Theoretical knowledge, expert assessment of likely behavior, and general understanding of the nature of the relationship are important considerations for choosing variables and functional form.
3. If an estimated equation has coefficients with unexpected signs, or unrealistic magnitudes, they could be caused by a misspecification such as the omission of an important variable.
4. Patterns in least squares residuals can be helpful for uncovering problems caused by an incorrect functional form. Some illustrations are given in Section 4.3.4.
5. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests. That is, t-tests for hypotheses such as $H_0 : \beta_3 = 0$ or F-tests for hypotheses such as $H_0 : \beta_3 = \beta_4 = 0$. Such tests can include coefficients of squares and products of variables as tests for a suitable functional form. Failure to reject a null hypotheses that one or more coefficients are zero can be an indication that the variable(s) are irrelevant. However, it is important to remember that failure to reject a null hypothesis can also occur if the data are not sufficiently rich to disprove the hypothesis. More will be said about poor data in Section 6.5. For the moment we note that, when

a variable has an insignificant coefficient, it can either be (a) discarded as an irrelevant variable, or (b) retained because the theoretical reason for its inclusion is a strong one.

6. Have the leverage, studentized residuals, DFBETAS, and DFFITS measures identified any influential observations?[7] If an unusual observation is not a data error, then understanding why it occurred may provide useful information for setting up the model.
7. Are the estimated coefficients robust with respect to alternative specifications? If the model is designed to be a causal one, and estimates of the causal coefficient change dramatically when different specifications of the model are estimated, or different sets of control variables are included, then there is cause for concern.
8. A test known as RESET (Regression Specification Error Test) can be useful for detecting omitted variables or an incorrect functional form. Details of this test are provided in Section 6.3.6.
9. Various model selection criteria, based on maximizing R^2, or minimizing the sum of squared errors (*SSE*), subject to a penalty for too many variables, have been suggested. These criteria are more valuable when a model is designed for prediction rather than causal analysis. For reliable prediction a sum of squared errors that is small relative to the explanatory power of the model is essential. We describe three of these criteria in Section 6.4.1: an adjusted R^2, the Akaike information criterion (AIC), and the Schwarz criterion (SC), also known as the Bayesian information criterion (BIC).
10. A more stringent assessment of a model's predictive ability is to use a "hold-out" sample. A least squares estimated equation is designed to minimize the within-sample sum of squared errors. To check out a model's ability to predict outside the sample, some observations can be withheld from estimation and the model can be assessed on its ability to predict the withheld observations. More details are provided in Section 6.4.1.
11. Following the guidelines in the previous 10 points can almost inevitably lead to revisions of originally proposed models, or to more general experimentation with alternative models. Searching for a model with "significant" estimates and the selective reporting of a finally chosen "significant" model is a questionable practice. Not knowing the search process that led to the selected results makes valid interpretation of the results difficult. Proper reporting of results should include disclosure of all estimated models and the criteria used for model selection.

6.3.6 RESET

Testing for model misspecification is a way of asking whether our model is adequate, or whether we can improve on it. It could be misspecified if we have omitted important variables, included irrelevant ones, chosen a wrong functional form, or have a model that violates the assumptions of the multiple regression model. **RESET** (REgression Specification Error Test) is designed to detect omitted variables and incorrect functional form. It proceeds as follows.

Suppose that we have specified and estimated the regression model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Let (b_1, b_2, b_3) be the least squares estimates, and let

$$\hat{y} = b_1 + b_2 x_2 + b_3 x_3 \tag{6.32}$$

be the fitted values of y. Consider the following two artificial models:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + e \tag{6.33}$$

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e \tag{6.34}$$

[7] These measures for detecting influential observations are discussed in Sections 4.3.6 and 6.5.3.

In (6.33), a test for misspecification is a test of $H_0: \gamma_1 = 0$ against the alternative $H_1: \gamma_1 \neq 0$. In (6.34), testing $H_0: \gamma_1 = \gamma_2 = 0$ against $H_1: \gamma_1 \neq 0$ and/or $\gamma_2 \neq 0$ is a test for misspecification. In the first case, a t- or an F-test can be used. An F-test is required for the second equation. Rejection of H_0 implies that the original model is inadequate and can be improved. A failure to reject H_0 says that the test has not been able to detect any misspecification.

To understand the idea behind the test, note that $\hat{y}^2$ and $\hat{y}^3$ will be polynomial functions of x_2 and x_3. If you square and cube both sides of (6.32), you will get terms such as $x_2^2, x_3^3, x_2x_3, x_2x_3^2$, and so on. Since polynomials can approximate many different kinds of functional forms, if the original functional form is not correct, the polynomial approximation that includes $\hat{y}^2$ and $\hat{y}^3$ may significantly improve the fit of the model. If it does, this fact will be detected through nonzero values of γ_1 and γ_2. Furthermore, if we have omitted variables and these variables are correlated with x_2 and x_3, then they are also likely to be correlated with terms such as x_2^2 and x_3^2, so some of their effect may be picked up by including the terms $\hat{y}^2$ and/or $\hat{y}^3$. Overall, the general philosophy of the test is if we can significantly improve the model by artificially including powers of the predictions of the model, then the original model must have been inadequate.

EXAMPLE 6.14 | Applying RESET to Family Income Equation

To illustrate RESET we return to the family income equation considered in Examples 6.10–6.12. In those examples specifications with different variables included were estimated, and the results presented in Table 6.1. The full model, without the irrelevant variables, was

$$\ln(FAMINC) = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + \beta_4 KL6 + e$$

Please go back and check Table 6.1, where RESET p-values for both $H_0: \gamma_1 = 0$ and $H_0: \gamma_1 = \gamma_2 = 0$ are presented in the last two rows of the table. The only instance where RESET rejects a model at a 5% significance level is where wife's education has been excluded and the null hypothesis is $H_0: \gamma_1 = \gamma_2 = 0$. Exclusion of *KL6* is not picked up by RESET, most likely because it is not highly correlated with *HEDU* and *WEDU*. Also, when the irrelevant variables *XTRA_X5* and *XTRA_X6* are included, and *WEDU* is excluded, RESET does not pick up the misspecification. The likely cause of this failure is the high correlations between *WEDU* and the two irrelevant variables.

There are two important lessons from this example. First, if RESET does not reject a model, that model is not necessarily a good one. Second, RESET will not always discriminate between alternative models. Rejection of the null hypothesis is indicative of a misspecification, but failure to reject the null hypothesis tells us very little.

6.4 Prediction

The prediction or forecasting problem for a regression model with one explanatory variable was introduced in Section 4.1. That material extends naturally to the more general model that has more than one explanatory variable. In this section, we describe that extension, reinforce earlier material, and provide some more general background.

Suppose we have values on $K-1$ explanatory variables represented by $\mathbf{x}_0 = (1, x_{02}, x_{03}, \ldots, x_{0K})$, and that we wish to use this information to predict or forecast a corresponding dependent variable value y_0. In Appendix 4D we learned that the minimum mean square error predictor for y_0 is the conditional expectation $E(y_0|\mathbf{x}_0)$. To make this result operational, we need to make an assumption about the functional form for $E(y_0|\mathbf{x}_0)$, and estimate the parameters on which it depends. In line with the multiple regression model, we assume that the conditional expectation is the linear-in-the-parameters function

$$E(y_0|\mathbf{x}_0) = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K} \tag{6.35}$$

Defining $e_0 = y_0 - E(y_0|\mathbf{x}_0)$, we can write

$$y_0 = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K} + e_0 \tag{6.36}$$

To estimate the parameters $(\beta_1, \beta_2, \ldots, \beta_K)$ in (6.35), we assume we have $i = 1, 2, \ldots, N$ observations y_i and $\mathbf{x}_i = (1, x_{i2}, x_{i3}, \ldots, x_{iK})$ such that

$$E(y_i|\mathbf{x}_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} \tag{6.37}$$

Define $e_i = y_i - E(y_i|\mathbf{x}_i)$ so that the model used to estimate $(\beta_1, \beta_2, \ldots, \beta_K)$ can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i \tag{6.38}$$

Equations (6.35)–(6.38) make up the **predictive model**. Equations (6.37) and (6.38) refer to the sample observations used to estimate the parameters. Equation (6.35) is the predictor that would be used if the parameters $(\beta_1, \beta_2, \ldots \beta_K)$ were known. Equation (6.36) incorporates the realized value y_0 and the error e_0. When we think of prediction or forecasting—we use the two terms interchangeably—we naturally think of forecasting outside the sample observations. Under these circumstances y_0 will be unobserved at the time the forecast is made. With time-series data, $\mathbf{x}_0$ will be future values of the explanatory variables for which a forecast is required; for cross-section data it will be values for an individual or some other economic unit that was not sampled. There are, however, instances where we make within-sample predictions or forecasts despite the fact that we have observed realized values for y for those observations. One example is their use in RESET where the regression equation was augmented with the squares and cubes of the within-sample predictions. When we are considering within-sample predictions, $\mathbf{x}_0$ will be identical to one of the $\mathbf{x}_i$, or it can be viewed as generic notation to represent all $\mathbf{x}_i$.

Note that (6.36) and (6.38) do **not** have to be causal models. To have a good predictive model, (y_i, y_0) needs to be highly correlated with the variables in $(\mathbf{x}_i, \mathbf{x}_0)$, but there is no requirement that (y_i, y_0) be caused by $(\mathbf{x}_i, \mathbf{x}_0)$. There is no requirement that all variables that affect y have to be included and there is no such thing as omitted variable bias. In (6.38), we are simply estimating the conditional expectation of the variables that are included. Under these circumstances, the interpretation of (e_i, e_0) is different from its interpretation in a **causal model**. In a causal model e represents the effect of variables omitted from the model; it is important that these effects are isolated from those in the model through the exogeneity assumption. We think of e as part of the data generating process. In a predictive model the coefficients in the conditional expectation can represent the direct effect of included variables and the indirect effect of excluded variables. The error term e is simply the difference between the realized value y and its conditional expectation; it is the **forecasting error** that would occur if $(\beta_1, \beta_2, \ldots, \beta_K)$ were known and did not have to be estimated. It does not take on an "all-other-variables" interpretation.

Application of least squares to (6.35) will yield unbiased estimates of $(\beta_1, \beta_2, \ldots, \beta_K)$ conditional on $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. If we assume further that $\text{var}(e_i|\mathbf{X}) = \sigma^2$ and $E(e_i e_j|\mathbf{X}) = 0$ for $i \neq j$, then the least squares estimator is best linear unbiased conditional on $\mathbf{X}$. Unconditionally, it will be a consistent estimator providing assumptions about the limiting behavior of the explanatory variables hold.[8] Having obtained the least squares estimates (b_1, b_2, …, b_K), we can define an operational predictor for y_0 as (6.35) with the unknown β_k replaced by their estimators. That is,

$$\hat{y}_0 = b_1 + b_2 x_{02} + b_3 x_{03} + \cdots + b_K x_{0K} \tag{6.39}$$

An extra assumption that we need is that $(e_0|\mathbf{x}_0)$ is uncorrelated with $(e_i|\mathbf{X})$ for $i = 1, 2, \ldots, N$ and $i \neq 0$. We also assume $\text{var}(e_0|\mathbf{x}_0) = \text{var}(e_i|\mathbf{X}) = \sigma^2$, an assumption used when deriving the variance of the forecast error.

After replacing the β_k with b_k, the forecast error is given by

$$\begin{aligned} f &= y_0 - \hat{y}_0 \\ &= (\beta_1 - b_1) + (\beta_2 - b_2) x_{02} + (\beta_3 - b_3) x_{03} + \cdots + (\beta_K - b_K) x_{0K} + e_0 \end{aligned} \tag{6.40}$$

There are two components in this forecast error: the errors $(\beta_k - b_k)$ from estimating the unknown parameters, and an error e_0 which is the deviation of the realized y_0 from its conditional mean. The predictor $\hat{y}_0$ is unbiased in the sense that $E(f|\mathbf{x}_0, \mathbf{X}) = 0$ and it is a best linear unbiased

[8] See Section 5.7.1 for an illustration in the case of simple regression.

predictor in the sense that the conditional variance $\text{var}(f|\mathbf{x}_0, \mathbf{X})$ is no greater than that of any other linear unbiased predictor. The conditional variance of the prediction error is

$$\begin{aligned}\text{var}(f|\mathbf{x}_0, \mathbf{X}) &= \text{var}\left[\left(\sum_{k=1}^{K}(\beta_k - b_k)x_{0k}\right)\middle|\mathbf{x}_0, \mathbf{X}\right] + \text{var}(e_0|\mathbf{x}_0, \mathbf{X}) \\ &= \text{var}\left[\left(\sum_{k=1}^{K} b_k x_{0k}\right)\middle|\mathbf{x}_0, \mathbf{X}\right] + \sigma^2 \\ &= \sum_{k=1}^{K} x_{0k}^2 \text{var}(b_k|\mathbf{x}_0, \mathbf{X}) + 2\sum_{k=1}^{K}\sum_{j=k+1}^{K} x_{0k}x_{0j}\text{cov}(b_k, b_j|\mathbf{x}_0, \mathbf{X}) + \sigma^2 \end{aligned} \tag{6.41}$$

In the first line of this equation we have assumed that the covariance between $(\beta_k - b_k)$ and e_0 is zero. This assumption will indeed be true for out-of-sample prediction and where e_0 is uncorrelated with the sample data used to estimate the β_k. For within-sample prediction the situation is more complicated. Strictly speaking, if e_0 is equal to one of the e_i in the sample, then $(\beta_k - b_k)$ and e_0 will be correlated. This correlation will not be large relative to the overall variance of f, however, and tends to get ignored in software calculations. In the second line of (6.41) $\beta_k x_{0k}$ can be treated as a constant and so $\text{var}\left((\beta_k - b_k)x_{0k}|\mathbf{x}_0, \mathbf{X}\right) = \text{var}(b_k x_{0k}|\mathbf{x}_0, \mathbf{X})$. The third line follows from the rule for calculating the variance of a weighted sum in (P.20) of the Probability Primer.

Each of the terms in the expression for $\text{var}(f|\mathbf{x}_0, \mathbf{X})$ involves σ^2. To obtain the estimated variance of the forecast error $\widehat{\text{var}}(f|\mathbf{x}_0, \mathbf{X})$, we replace σ^2 with its estimator $\hat{\sigma}^2$. The standard error of the forecast is given by $\text{se}(f) = \sqrt{\widehat{\text{var}}(f|\mathbf{x}_0, \mathbf{X})}$. If the random errors e_i and e_0 are normally distributed, or if the sample is large, then

$$\frac{f}{\text{se}(f)} = \frac{y_0 - \hat{y}_0}{\sqrt{\widehat{\text{var}}(y_0 - \hat{y}_0|\mathbf{x}_0, \mathbf{X})}} \sim t_{(N-K)} \tag{6.42}$$

Following the steps we have used many times, a $100(1-\alpha)\%$ interval predictor for y_0 is $\hat{y}_0 \pm t_c \text{se}(f)$, where t_c is a critical value from the $t_{(N-K)}$-distribution.

Before providing an example there are two practical considerations worth mentioning. First, in (6.41), the error variance σ^2 is typically much larger than the variance of the other component—that part of the forecast error attributable to estimation of the β_k. Consequently, this latter component is sometimes ignored and $\text{se}(f) = \hat{\sigma}$ is used. Second, the framework presented so far does not capture many of the typical characteristics of time-series forecasting. With time-series forecasting, some of the explanatory variables will usually be lagged values of the dependent variable. This means that the conditional expectation of a y_0 will depend on past values of itself. The sample information contributes to the conditional expectation of y_0. In the above exposition we have treated $\mathbf{x}_0$ as future values of the explanatory variables. The sample information has only contributed to the predictor through the estimation of the unknown β_k. In other words, $E(y_0|\mathbf{x}_0) = E(y_0|\mathbf{x}_0, \mathbf{X}, \mathbf{y})$, where $\mathbf{y}$ is used to denote all observations on the dependent variable. A more general scenario for time-series forecasting where this assumption is relaxed is considered in Chapter 9.

EXAMPLE 6.15 | Forecasting *SALES* for the Burger Barn

We are concerned with finding a 95% prediction interval for *SALES* at Big Andy's Burger Barn when $PRICE_0 = 6$, $ADVERT_0 = 1.9$ and $ADVERT_0^2 = 3.61$. These are the values considered by Big Andy in Example 6.6. In terms of the general notation $\mathbf{x}_0 = (1, 6, 1.9, 3.61)$. The point prediction is

$$\begin{aligned}\widehat{SALES}_0 &= 109.719 - 7.640\,PRICE_0 + 12.1512\,ADVERT_0 \\ &\quad - 2.768\,ADVERT_0^2 \\ &= 109.719 - 7.640 \times 6 + 12.1512 \times 1.9 - 2.768 \\ &\quad \times 3.61 \\ &= 76.974\end{aligned}$$

With the settings proposed by Big Andy, we forecast that sales will be \$76,974.

To obtain a prediction interval, we first need to compute the estimated variance of the forecast error. Using equation (6.41) and the covariance matrix values in Table 6.3, we have

$$\begin{aligned}
\widehat{\text{var}}(f|\mathbf{x}_0,\mathbf{X}) &= \hat{\sigma}^2 + \widehat{\text{var}}(b_1|\mathbf{x}_0,\mathbf{X}) + x_{02}^2\widehat{\text{var}}(b_2|\mathbf{x}_0,\mathbf{X}) \\
&\quad + x_{03}^2\widehat{\text{var}}(b_3|\mathbf{x}_0,\mathbf{X}) + x_{04}^2\widehat{\text{var}}(b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}\widehat{\text{cov}}(b_1,b_2|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{03}\widehat{\text{cov}}(b_1,b_3|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{04}\widehat{\text{cov}}(b_1,b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}x_{03}\widehat{\text{cov}}(b_2,b_3|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}x_{04}\widehat{\text{cov}}(b_2,b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{03}x_{04}\widehat{\text{cov}}(b_3,b_4|\mathbf{x}_0,\mathbf{X}) \\
&= 21.57865 + 46.22702 + 6^2 \times 1.093988 \\
&\quad + 1.9^2 \times 12.6463 + 3.61^2 \times 0.884774 \\
&\quad + 2 \times 6 \times (-6.426113) \\
&\quad + 2 \times 1.9 \times (-11.60096) \\
&\quad + 2 \times 3.61 \times 2.939026 \\
&\quad + 2 \times 6 \times 1.9 \times 0.300406 \\
&\quad + 2 \times 6 \times 3.61 \times (-0.085619) \\
&\quad + 2 \times 1.9 \times 3.61 \times (-3.288746) \\
&= 22.4208
\end{aligned}$$

TABLE 6.3 Covariance Matrix for Andy's Burger Barn Model

	b_1	b_2	b_3	b_4
b_1	46.227019	−6.426113	−11.600960	2.939026
b_2	−6.426113	1.093988	0.300406	−0.085619
b_3	−11.600960	0.300406	12.646302	−3.288746
b_4	2.939026	−0.085619	−3.288746	0.884774

The standard error of the forecast error is $\text{se}(f) = \sqrt{22.4208} = 4.7351$, and the relevant t-value is $t_{(0.975,\,71)} = 1.9939$, giving a 95% prediction interval of

$$\begin{aligned}
&(76.974 - 1.9939 \times 4.7351,\ \ 76.974 + 1.9939 \times 4.7351) \\
&\quad = (67.533,\ \ 86.415)
\end{aligned}$$

We predict, with 95% confidence, that Big Andy's settings for price and advertising expenditure will yield *SALES* between \$67,533 and \$86,415.

6.4.1 Predictive Model Selection Criteria

In this section we consider three model selection criteria: (i) R^2 and $\overline{R}^2$, (ii) AIC, and (iii) SC (BIC), and describe how a hold-out sample can be used to evaluate a model's predictive or forecast ability. Throughout the section you should keep in mind that we are not recommending blind application of any of these criteria. They should be treated as devices that provide additional information about the relative merits of alternative models, and they should be used in conjunction with the other considerations listed in Section 6.3.5 and in the introduction to Section 6.3.

Choice of a model based exclusively on $\overline{R}^2$, AIC, or SC involves choosing a model that minimizes the sum of squared errors with a penalty for adding extra variables. While these criteria can be used for both predictive and causal models, their goal of minimizing a function of the sum of squared errors rather than focusing on the coefficient, make them more suitable for predictive model selection. Another common feature of the criteria is that they are suitable only for comparing models with the same dependent variable, not for models with different dependent variables such as y and $\ln(y)$. More general versions of the AIC and SC, based on likelihood functions[9],

[9] An introduction to maximum likelihood estimation can be found in Appendix C.8.

are available for models with transformations of the dependent variable, but we do not consider them here.

R^2 and $\bar{R}^2$ In Chapters 4 and 5, we introduced the coefficient of determination $R^2 = 1 - SSE/SST$ as a measure of goodness of fit. It shows the proportion of variation in a dependent variable explained by variation in the explanatory variables. Since it is desirable to have a model that fits the data well, there can be a tendency to think that the best model is the one with the highest R^2. There are at least two problems with this line of thinking. First, if cross-sectional data are being used to estimate a causal effect, then low R^2's are typical and not necessarily a concern. What is more important is to avoid omitted variable bias and to have a sample size sufficiently large to get a reliable estimate of the coefficient of interest.

The second problem is one related to predictive models, namely, that comparing models on the basis of R^2 is only legitimate if the models have the same number of explanatory variables. Adding more variables always increases R^2 even if the variables added have no justification. As variables are added the sum of squared errors SSE goes down and thus R^2 goes up. If the model contains $N - 1$ variables, then $R^2 = 1$.

An alternative measure of goodness of fit called the **adjusted-R^2**, denoted as $\bar{R}^2$, has been suggested to overcome this problem. It is computed as

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

This measure does not always go up when a variable is added because of the degrees of freedom term $N - K$ in the numerator. As the number of variables K increases, SSE goes down, but so does $N - K$. The effect on $\bar{R}^2$ depends on the amount by which SSE falls. While solving one problem, this corrected measure of goodness of fit unfortunately introduces other problems. It loses its interpretation; $\bar{R}^2$ is no longer the proportion of explained variation. Also, it can be shown that if a variable is added to an equation, say with coefficient β_K, then $\bar{R}^2$ will increase if the t-value for testing the hypothesis $H_0 : \beta_K = 0$ is greater than one. Thus, using $\bar{R}^2$ as a device for selecting the appropriate set of explanatory variables is like using a hypothesis test for significance of a coefficient with a critical value of 1, a value much less than that typically used with 5% and 10% levels of significance. Because of these complications, we prefer to report the unadjusted R^2 as a goodness-of-fit measure, and caution is required if $\bar{R}^2$ is used for model selection. Nevertheless, you should be familiar with $\bar{R}^2$. You will see it in research reports and on the output of software packages.

Information Criteria Selecting variables to maximize $\bar{R}^2$ can be viewed as selecting variables to minimize SSE, subject to a penalty for introducing too many variables. Both the AIC and the SC work in a similar way but with different penalties for introducing too many variables. The **Akaike information criterion (AIC)** is given by

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N} \tag{6.43}$$

and the **Schwarz criterion (SC)**, also known as the **Bayesian information criterion** (**BIC**), is given by

$$\text{SC} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N} \tag{6.44}$$

In each case the first term becomes smaller as extra variables are added, reflecting the decline in the SSE, but the second term becomes larger because K increases. Because $K\ln(N)/N > 2K/N$

for $N \geq 8$, in reasonable sample sizes the SC penalizes extra variables more heavily than does the AIC. Using these criteria, the model with the smallest AIC, or the smallest SC, is preferred.

To get values of the more general versions of these criteria based on maximized values of the likelihood function you need to add $[1 + \ln(2\pi)]$ to (6.43) and (6.44). It is good to be aware of this fact in case your computer software reports the more general versions. However, although it obviously changes the AIC and SC values, adding a constant does not change the choice of variables that minimize the criteria.

Using a Hold-Out Sample

When a model is designed for prediction or forecasting, we are naturally interested in its ability to forecast dependent variable values that have not yet been observed. To assess a model on this basis, we could make some forecasts and then compare these forecasts with the corresponding realizations after they occur. However, if we are in the model construction phase of an investigation, it is unlikely we would want to wait for extra observations. A way out of this dilemma is to hold back some of the observations from estimation and then evaluate the model on the basis of how well it can predict the omitted observations. Suppose we have a total of N observations of which N_1 are used for estimation and $N_2 = N - N_1$ are held back to evaluate a model's forecasting ability. Thus, we have estimates $(b_1, b_2, \ldots, b_K)$ from observations $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N_1$ and we calculate the predictions

$$\hat{y}_i = b_1 + b_2 x_{i2} + \cdots + b_K x_{iK} \quad i = N_1 + 1, N_2 + 2, \ldots, N$$

A measure of the model's out-of-sample forecasting ability is the **root mean squared error** (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N_2} \sum_{i=N_1+1}^{N} (y_i - \hat{y}_i)^2}$$

We expect this quantity to be larger than its within-sample counterpart $\hat{\sigma} = \sqrt{\sum_{i=1}^{N_1} (y_i - \hat{y}_i)^2 / (N_1 - K)}$ because the least squares estimation procedure is such that $\sum_{i=1}^{N_1} (y_i - \hat{y}_i)^2$ is minimized. Models can be compared on the basis of their hold-out RMSEs.

EXAMPLE 6.16 | Predicting House Prices

Real estate agents and potential homebuyers are interested in valuing houses or predicting the price of a house with particular characteristics. There are many factors that have a bearing on the price of a house, but for our predictive model we will consider just two, the age of the house in years (*AGE*), and its size in hundreds of square feet (*SQFT*). The most general model we consider is

$$\ln(PRICE) = \beta_1 + \beta_2 AGE + \beta_3 SQFT + \beta_4 AGE^2 + \beta_5 SQFT^2 + \beta_6 AGE \times SQFT + e$$

where *PRICE* is the house price in thousands of dollars. Of interest is whether some or all of the quadratic terms AGE^2, $SQFT^2$, and $AGE \times SQFT$ improve the predictive ability of the model. For convenience, we evaluate predictive ability in terms of ln(*PRICE*) not *PRICE*. We use data on 900 houses sold in Baton Rouge, Louisiana in 2005, stored in the data file *br5*. For a comparison based on the RMSE of predictions (but not the other criteria) we randomly chose 800 observations for estimation and 100 observations for the hold-out sample. After this random selection, the observations were ordered so that the first 800 were used for estimation and the last 100 for predictive assessment.

Values of the criteria for the various models appear in Table 6.4. Looking for the model with the highest $\bar{R}^2$, and the models with the smallest values (or largest negative numbers) for the AIC and SC, we find that all three criteria prefer model 2 where AGE^2 is included, but $SQFT^2$ and $AGE \times SQFT$ are excluded. Using the out-of-sample RMSE criterion, model 6, with $AGE \times SQFT$ included in addition to AGE^2, is slightly favored over model 2.

TABLE 6.4 Model Selection Criteria for House Price Example

Model	Variables included in addition to (*SQFT*, *AGE*)	R^2	$\bar{R}^2$	AIC	SC	RMSE
1	None	0.6985	0.6978	–2.534	–2.518	0.2791
2	AGE^2	0.7207	0.7198*	–2.609*	–2.587*	0.2714
3	$SQFT^2$	0.6992	0.6982	–2.535	–2.513	0.2841
4	$AGE \times SQFT$	0.6996	0.6986	–2.536	–2.515	0.2790
5	AGE^2, $SQFT^2$	0.7208	0.7196	–2.607	–2.580	0.2754
6	AGE^2, $AGE \times SQFT$	0.7210	0.7197	–2.608	–2.581	0.2712*
7	$SQFT^2$, $AGE \times SQFT$	0.7006	0.6993	–2.537	–2.510	0.2840
8	$SQFT^2$, AGE^2, $AGE \times SQFT$	0.7212*	0.7197	–2.606	–2.574	0.2754

*Best model according to each of the criteria.

6.5 Poor Data, Collinearity, and Insignificance

Most economic data that are used for estimating economic relationships are nonexperimental. Indeed, in most cases they are simply "collected" for administrative or other purposes. They are not the result of a planned experiment in which an experimental design is specified for the explanatory variables. In controlled experiments the right-hand-side variables in the model can be assigned values in such a way that their individual effects can be identified and estimated with precision. When data are the result of an uncontrolled experiment, many of the economic variables may move together in systematic ways. Such variables are said to be **collinear**, and the problem is labeled **collinearity**. In this case there is neither a guarantee that the data will be "rich in information" nor that it will be possible to isolate the economic relationship or parameters of interest.

As an example, consider the problem faced by the marketing executives at Big Andy's Burger Barn when they try to estimate the increase in sales revenue attributable to advertising that appears in newspapers *and* the increase in sales revenue attributable to coupon advertising. Suppose that it has been common practice to coordinate these two advertising devices, so that at the same time that advertising appears in the newspapers there are flyers distributed containing coupons for price reductions on hamburgers. If variables measuring the expenditures on these two forms of advertising appear on the right-hand side of a sales revenue equation such as (5.2), then the data on these variables will show a systematic, positive relationship; intuitively, it will be difficult for such data to reveal the separate effects of the two types of ads. Although it is clear that total advertising expenditure increases sales revenue, because the two types of advertising expenditure move together, it may be difficult to sort out their separate effects on sales revenue.

As a second example, consider a production relationship explaining output over time as a function of the amounts of various quantities of inputs employed. There are certain factors of production (inputs), such as labor and capital, that are used in *relatively fixed proportions*. As production increases, the changing amounts of two or more such inputs reflect equiproportionate increases. Proportional relationships between variables are the very sort of systematic relationships that epitomize "collinearity." Any effort to measure the individual or separate effects (marginal products) of various mixes of inputs from such data will be difficult.

It is not just relationships between variables in a sample of data that make it difficult to isolate the separate effects of individual explanatory variables. If the values of an explanatory variable

do not vary or change much within a sample of data, then it is clearly difficult to use that data to estimate a coefficient that describes the effect of change in that variable. It is hard to estimate the effect of change if there has been no change.

6.5.1 The Consequences of Collinearity

The consequences of collinearity and/or lack of variation depend on whether we are examining an extreme case in which estimation breaks down or a bad, but not extreme, case in which estimation can still proceed but our estimates lack precision. In Section 5.3.1, we considered the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

and wrote the variance of the least squares estimator for β_2 as

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{(1 - r_{23}^2)\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2} \tag{6.45}$$

where r_{23} is the correlation between x_2 and x_3. Exact or extreme collinearity exists when x_2 and x_3 are perfectly correlated, in which case $r_{23} = 1$ and $\text{var}(b_2|\mathbf{X})$ goes to infinity. Similarly, if x_2 exhibits no variation $\sum(x_{i2} - \bar{x}_2)^2$ equals zero and $\text{var}(b_2|\mathbf{X})$ again goes to infinity. In this case, x_2 is collinear with the constant term. In general, w*henever there are one or more* ***exact*** *linear relationships among the explanatory variables, then the condition of exact collinearity exists. In this case, the least squares estimator is not defined*. We *cannot* obtain estimates of β_k's using the least squares principle. One of our least squares assumptions MR5, which says that the values of x_{ik} are not exact linear functions of the other explanatory variables, is violated.

The more usual case is one in which correlations between explanatory variables might be high, but not exactly one; variation in explanatory variables may be low but not zero; or linear dependencies between more than two explanatory variables could be high but not exact. These circumstances do *not* constitute a violation of least squares assumptions. By the Gauss–Markov theorem, the least squares estimator is still the best linear unbiased estimator. We might still be unhappy, however, if the best we can do is constrained by the poor characteristics of our data. From (6.45), we can see that when r_{23} is close to one or $\sum(x_i - \bar{x}_2)^2$ is close to zero, the variance of b_2 will be large. A large variance means a large standard error, which means the estimate may not be significantly different from zero and an interval estimate will be wide. The sample data have provided relatively imprecise information about the unknown parameters.

Although (6.45) is only valid for a regression model with two explanatory variables, with a few simple changes we can generalize this equation to gain insights into collinearity in the more general multiple regression model with $K - 1$ explanatory variables. First, recall from Section 4.2.2 that a simple correlation between two variables is the same as the R^2 from the regression of one variable on another, so that $r_{23}^2 = R_{2.}^2$, where $R_{2.}^2$ is the R^2 from the so-called **auxiliary regression** $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + v_i$. Then, another way to write (6.45) is

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{\sum(x_{i2} - \bar{x}_2)^2 (1 - R_{2.}^2)} \tag{6.46}$$

The beauty of this equation is that it holds for the general model $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$, where $R_{2.}^2$ is the R^2 from the auxiliary regression $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + \cdots + \alpha_K x_{iK} + v_i$. The ratio

$$\text{VIF} = 1/(1 - R_{2.}^2)$$

is called the **variance inflation factor**. If $R_{2.}^2 = 0$, indicating no collinearity—no variation in x_2 can be explained by the other explanatory variables—then VIF = 1 and $\text{var}(b_2|\mathbf{X}) = \sigma^2/\sum(x_{i2} - \bar{x}_2)^2$. On the other hand, if $R_{2.}^2 = 0.90$, indicating that 90% of the variation in x_2 can be explained by the other regressors, then VIF = 10 and $\text{var}(b_2|\mathbf{X})$ is ten times

larger than the than it would be if there was no collinearity present. VIF is sometimes used to describe the severity of collinearity in a regression. Auxiliary regression R_k^2's and variance inflation factors can be found for every explanatory variable in a regression; equations analogous to (6.46) hold for each of the coefficient estimates.

By examining $R_{2.}^2$, we can obtain a very informative third representation. The R^2 from the regression $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + \cdots + \alpha_K x_{iK} + v_i$ is the portion of the total variation in x_2 about its mean, $\sum(x_{i2} - \bar{x}_2)^2$, explained by the model. Let the fitted values from the auxiliary regression be $\hat{x}_{i2} = a_2 + a_3 x_{i3} + \cdots + a_K x_{iK}$, where $(a_2, a_3, \ldots, a_K)$ are the least squares estimates of $(\alpha_2, \alpha_3, \ldots, \alpha_K)$. A residual from the auxiliary regression is $x_{i2} - \hat{x}_{i2}$ and its R^2 can be written as

$$R_{2.}^2 = 1 - \frac{\sum(x_{i2} - \hat{x}_{i2})^2}{\sum(x_{i2} - \bar{x}_2)^2}$$

Substituting this into (6.46), we have

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{\sum(x_{i2} - \hat{x}_{i2})^2} \tag{6.47}$$

The term $\sum(x_{i2} - \hat{x}_{i2})^2$ is the sum of squared least squares residuals from the auxiliary regression. When collinearity is stronger, with a larger amount of variation in x_2 explained by the other regressors, the smaller $\sum(x_{i2} - \hat{x}_{i2})^2$ becomes and the larger $\text{var}(b_2|\mathbf{X})$ becomes. It is the variation in x_2 that is *not* explained by the other regressors that increases the precision of least squares estimation.

The effects of imprecise estimation resulting from collinearity can be summarized as follows:

1. When estimator standard errors are large, it is likely that the usual t-tests will lead to the conclusion that parameter estimates are not significantly different from zero. This outcome occurs despite possibly high R^2- or F-values indicating significant explanatory power of the model as a whole. The problem is that collinear variables do not provide enough information to estimate their separate effects, even though theory may indicate their importance in the relationship.
2. Estimators may be very sensitive to the addition or deletion of a few observations, or to the deletion of an apparently insignificant variable.
3. Despite the difficulties in isolating the effects of individual variables from such a sample, accurate forecasts may still be possible if the nature of the collinear relationship remains the same within the out-of-sample observations. For example, in an aggregate production function where the inputs labor and capital are nearly collinear, accurate forecasts of output may be possible for a particular ratio of inputs but not for various mixes of inputs.

6.5.2 Identifying and Mitigating Collinearity

Because nonexact collinearity is not a violation of least squares assumptions, it does not make sense to go looking for a problem if there is no evidence that one exists. If you have estimated an equation where the coefficients are precisely estimated and significant, they have the expected signs and magnitudes, and they are not sensitive to adding or deleting a few observations, or an insignificant variable, then there is no reason to try and identify or mitigate collinearity. If there are highly correlated variables, they are not causing you a problem. However, if you have a poorly estimated equation that does not live up to expectations, it is useful to establish why the estimates are poor.

One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables. These sample correlations are descriptive measures of linear association. However, collinear relationships that involve more than two explanatory

variables are better detected using **auxiliary regressions**. If an R_k^2 is high, say greater than 0.8, then a large portion of the variation in x_k is explained by the other regressors, and that may have a detrimental effect on the precision of estimation of β_k. If an auxiliary regression R_k^2 is not high, then the precision of an estimator b_k is not unduly affected by collinearity, although it may still suffer if the variation in x_k is inadequate.

The collinearity problem is that the data do not contain enough "information" about the individual effects of explanatory variables to permit us to estimate all the parameters of the statistical model precisely. Consequently, one solution is to obtain more information and include it in the analysis. One form the new information can take is more, and better, sample data. Unfortunately, in economics, this is not always possible. Cross-sectional data are expensive to obtain, and, with time-series data, one must wait for the data to appear. Alternatively, if new data are obtained via the same nonexperimental process as the original sample of data, then the new observations may suffer the same collinear relationships and provide little in the way of new, independent information. Under these circumstances the new data will help little to improve the precision of the least squares estimates.

A second way of adding new information is to introduce, as we did in Section 6.2, *nonsample* information in the form of restrictions on the parameters. This nonsample information may then be combined with the sample information to provide restricted least squares estimates. The good news is that using nonsample information in the form of linear constraints on the parameter values reduces estimator sampling variability. The bad news is that the resulting restricted estimator is *biased* unless the restrictions are *exactly* true. Thus it is important to use good nonsample information, so that the reduced sampling variability is not bought at a price of large estimator biases.

EXAMPLE 6.17 | Collinearity in a Rice Production Function

To illustrate collinearity we use data on rice production from a cross section of Philippine rice farmers to estimate the production function

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA) + \beta_3 \ln(LABOR) + \beta_4 \ln(FERT) + e \tag{6.48}$$

where *PROD* denotes tonnes of freshly threshed rice, *AREA* denotes hectares planted, *LABOR* denotes person-days of hired and family labor and *FERT* denotes kilograms of fertilizer. Data for the years 1993 and 1994 can be found in the file *rice5*. One would expect collinearity may be an issue. Larger farms with more area are likely to use more labor and more fertilizer than smaller farms. The likelihood of a collinearity problem is confirmed by examining the results in Table 6.5, where we have estimated the function using data from 1994 only. These results convey very little information. The 95% interval estimates are very wide, and, because the coefficients of ln(*AREA*) and ln(*LABOR*) are not significantly different from zero, their interval estimates include a negative range. The high auxiliary R^2's and correspondingly high variance inflation factors point to collinearity as the culprit for the imprecise results. Further evidence is a relatively high $R^2 = 0.875$ from estimating (6.48), and a p-value of 0.0021 for the joint test of the two insignificant coefficients, $H_0: \beta_2 = \beta_3 = 0$.

TABLE 6.5 **Rice Production Function Results from 1994 Data**

Variable	Coefficient b_k	se(b_k)	95% Interval Estimate	p-Value*	Auxiliary Regression R^2	VIF
C	−1.9473	0.7385		0.0119		
ln(*AREA*)	0.2106	0.1821	[−0.1573, 0.5786]	0.2543	0.891	9.2
ln(*LABOR*)	0.3776	0.2551	[−0.1379, 0.8931]	0.1466	0.944	17.9
ln(*FERT*)	0.3433	0.1280	[0.0846, 0.6020]	0.0106	0.870	7.7

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

We consider two ways of improving the precision of our estimates: (1) including non-sample information, and (2) using more observations. For non-sample information, suppose that we are willing to accept the notion of constant returns to scale. That is, increasing all inputs by the same proportion will lead to an increase in production of the same proportion. If this constraint holds, then $\beta_2 + \beta_3 + \beta_4 = 1$. Testing this constraint as a null hypothesis yields a p-value of 0.313; so it is not a constraint that is incompatible with the 1994 data. Substituting $\beta_2 + \beta_3 + \beta_4 = 1$ into (6.48) and rearranging the equation gives

$$\ln\left(\frac{PROD}{AREA}\right) = \beta_1 + \beta_3 \ln\left(\frac{LABOR}{AREA}\right) + \beta_4 \ln\left(\frac{FERT}{AREA}\right) + e \tag{6.49}$$

This equation can be viewed as a "yield" equation. Rice yield per hectare is a function of labor per hectare and fertilizer per hectare. Results from estimating it appear in Table 6.6. Has there been any improvement? The answer is not much! The estimate for β_3 is no longer "insignificant," but that is more attributable to an increase in the magnitude of b_3 than to a reduction in its standard error. The reduction in standard errors is only marginal, and the interval estimates are still wide, conveying little information. The squared correlation between $\ln(LABOR/AREA)$ and $\ln(FERT/AREA)$ is 0.414 which is much less than the earlier auxiliary R^2's, but, nevertheless, the new estimates are relatively imprecise.

As an alternative to injecting non-sample information into the estimation procedure, we examine the effect of including more observations by combining the 1994 data with observations from 1993. The results are given in Table 6.7. Here there has been a substantial reduction in the standard errors, with considerable improvement in the precision of the estimates, despite the fact that the variance inflation factors still remain relatively large. The greatest improvement has been for the coefficient of $\ln(FERT)$, which has the lowest variance inflation factor. The interval estimates for the other two coefficients are still likely to be wider than a researcher would desire, but at least there has been some improvement.

TABLE 6.6 **Rice Production Function Results from 1994 Data with Constant Returns to Scale**

Variable	Coefficient b_k	se(b_k)	95% Interval Estimate	p-Value*
C	−2.1683	0.7065		0.0038
$\ln(AREA)$	0.2262	0.1815	[−0.1474, 0.5928]	0.2197
$\ln(LABOR)$	0.4834	0.2332	[0.0125, 0.9544]	0.0445
$\ln(FERT)$	0.2904	0.1171	[0.0539, 0.5268]	0.0173

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

TABLE 6.7 **Rice Production Function Results from Data for 1993 and 1994**

Variable	Coefficient	se(b_k)	95% Interval Estimate	p-Value*	Auxiliary Regression R^2	VIF
C	−1.8694	0.4565		0.0001		
$\ln(AREA)$	0.2108	0.1083	[−0.0045, 0.4261]	0.0549	0.870	7.7
$\ln(LABOR)$	0.3997	0.1307	[0.1399, 0.6595]	0.0030	0.901	10.1
$\ln(FERT)$	0.3195	0.0635	[0.1932, 0.4457]	0.0000	0.776	4.5

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

TABLE 6.8 Statistics for Identifying Influential Observations

Influence Statistic	Formula	Investigative Threshold
Leverage	$h_i = \dfrac{\widehat{\text{var}}(\hat{e}_i) - \hat{\sigma}^2}{\hat{\sigma}^2}$	$h_i > \dfrac{2K}{N}$ or $\dfrac{3K}{N}$
Studentized residual	$\hat{e}_i^{stu} = \dfrac{\hat{e}_i}{\hat{\sigma}(i)(1-h_i)^{1/2}}$	$\left\|\hat{e}_i^{stu}\right\| > 2$
DFBETAS	$\text{DFBETAS}_{ki} = \dfrac{b_k - b_k(i)}{(\hat{\sigma}(i)/\hat{\sigma}) \times \text{se}(b_k)}$	$\left\|\text{DFBETAS}_{ki}\right\| > \dfrac{2}{\sqrt{N}}$
DFFITS	$\text{DFFITS}_i = \left(\dfrac{h_i}{1-h_i}\right)^{1/2} \hat{e}_i^{stu}$	$\left\|\text{DFFITS}_i\right\| > 2\left(\dfrac{K}{N}\right)^{1/2}$

6.5.3 Investigating Influential Observations

In Section 4.3.6, we introduced a number of measures for detecting influential observations. The purpose of having such measures is first to detect whether there may have been a data error, and second, if the accuracy of the data is confirmed, to identify unusual observations that may be worthy of further investigation. Are there observations that can be explained within the context of the proposed model? Are there other factors at work that could have led to the unusual observations?

In Section 4.3.6, the measures were introduced within the context of the simple regression model with one explanatory variable. The same measures are relevant for the multiple regression model, but some of the formulas change slightly to accommodate the extra regressors. Now would be a good time to go back and reread Section 4.3.6. Are you back? Now that you understand the concepts, we can proceed. The important concepts introduced in that section were the leverage of the ith observation, h_i, the studentized residual, $\hat{e}_i^{stu}$, the sensitivity of a coefficient estimate to omission of the ith observation, DFBETAS_{ki}, and the sensitivity of a prediction to omission of the ith observation DFFITS_i. Multiple regression versions of these measures are summarized in Table 6.8 along with conventional thresholds above which further scrutiny of an observation may be warranted. Remember, the purpose is not to throw out unusual observations but to learn from them. They may reveal some important characteristics of the data.

EXAMPLE 6.18 | Influential Observations in the House Price Equation

To illustrate the identification of potentially influential observations, we return to Example 6.16 where, using predictive model selection criteria, the preferred equation for predicting house prices was

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

In a sample of 900 observations it is not surprising to find a relatively large number of data points where the various influence measures exceed the recommended thresholds. As examples, in Table 6.9 we report the values of the measures for those observations with the three largest DFFITS. It turns out that the other influence measures for these three observations also have large values. In parentheses next to each of the values is the rank of its absolute value. When we check the characteristics of the three unusual observations, we find observation 540 is the newest house in the sample and observation 150 is the oldest house. Observation 411 is both old and large; it is the 10th largest (99th percentile) and the sixth oldest (percentile 99.4) house in the sample. In Exercise 6.20, you are invited to explore further the effect of these observations.

TABLE 6.9 Influence Measures for House Price Equation

Observation	h_i (rank)	$\hat{e}_i^{stu}$ (rank)	DFFITS$_i$ (rank)	DFBETAS$_{ki}$ (rank)		
Threshold	$\frac{2.5K}{N} = 0.011$	2	$2\left(\frac{K}{N}\right)^{1/2} = 0.133$	$\frac{2}{\sqrt{N}} = 0.067$		
				SQFT	*AGE*	*AGE*^{2}
411	0.0319 (10)	−4.98 (1)	0.904 (1)	−0.658 (1)	0.106 (17)	−0.327 (3)
524	0.0166 (22)	−4.31 (3)	0.560 (2)	0.174 (9)	0.230 (2)	−0.381 (2)
150	0.0637 (2)	1.96 (48)	−0.511 (3)	−0.085 (29)	−0.332 (1)	0.448 (1)

6.6 Nonlinear Least Squares

We have discovered how the least squares estimation technique can be used to estimate a variety of nonlinear functions. They include log-log models, log-linear models, and models with quadratic and interaction terms. However, the models we have encountered so far have all been linear in the parameters $\beta_1, \beta_2, \ldots, \beta_K$.[10] In this section we discuss estimation of models that are nonlinear in the parameters. To give an appreciation of what is meant by such a model, it is convenient to begin with the following simple artificial example,

$$y_i = \beta x_{i1} + \beta^2 x_{i2} + e_i \tag{6.50}$$

where y_i is a dependent variable, x_{i1} and x_{i2} are explanatory variables, β is an unknown parameter that we wish to estimate, and the e_i satisfy the multiple regression assumptions MR1–MR5. This example differs from the conventional linear model because the coefficient of x_{i2} is equal to the square of the coefficient of x_{i1}, and the number of parameters is not equal to the number of variables.

How can β be estimated? Think back to Chapter 2. What did we do when we had a simple linear regression equation with two unknown parameters β_1 and β_2? We set up a sum of squared errors function that, in the context of (6.50), is

$$S(\beta) = \sum_{i=1}^{N} \left(y_i - \beta x_{i1} - \beta^2 x_{i2}\right)^2 \tag{6.51}$$

Then we asked what values of the unknown parameters make $S(\beta)$ a minimum. We searched for the bottom of the bowl in Figure 2A.1. We found that we could derive formulas for the minimizing values b_1 and b_2. We called these formulas the least squares estimators.

When we have models that are nonlinear in the parameters, we cannot in general derive formulas for the parameter values that minimize the sum of squared errors function. However, for a given set of data, we can ask the computer to search for the parameter values that take us to the bottom of the bowl. There are many numerical software algorithms that can be used to find minimizing values for functions such as $S(\beta)$. Those minimizing values are known as the **nonlinear least squares estimates**. It is also possible to obtain numerical standard errors that assess the reliability of the nonlinear least squares estimates. Finite sample properties

[10] There have been a few exceptions where we have used notation other than $\beta_1, \beta_2, \ldots, \beta_K$ to denote the parameters.

and distributions of nonlinear least squares estimators are not available, but their large sample asymptotic properties are well established.[11]

EXAMPLE 6.19 | Nonlinear Least Squares Estimates for Simple Model

To illustrate estimation of (6.50), we use data stored in the file *nlls*. The sum of squared error function is graphed in Figure 6.1. Because we only have one parameter, we have a two-dimensional curve, not a "bowl." It is clear from the curve that the minimizing value for β lies between 1.0 and 1.5. From your favorite software, the nonlinear least squares estimate turns out to be $b = 1.1612$. The standard error depends on the degree of curvature of the sum of squares function at its minimum. A sharp minimum with a high degree of curvature leads to a relatively small standard error, while a flat minimum with a low degree of curvature leads to a relatively high standard error. There are different ways of measuring the curvature that can lead to different standard errors. In this example, the "outer-product of gradient" method yields a standard error of $\text{se}(b) = 0.1307$, while the standard error from the "observed-Hessian" method is $\text{se}(b) = 0.1324$.[12] Differences such as this one disappear as the sample size gets larger.

Two words of warning must be considered when estimating a nonlinear-in-the-parameters model. The first is to check that the estimation process has converged to a global minimum. The estimation process is an iterative one where a series of different parameter values are checked until the process converges at the minimum. If your software tells you the process has failed to converge, the output provided, if any, **does not** provide the nonlinear least squares estimates. This might happen if a maximum number of iterations has been reached or there has been a numerical problem that has caused the iterations to stop. A second problem that can occur is that the iterative process may stop at a "local" minimum rather than the "global" minimum. In the example in Figure 6.1, there is a local minimum at $\beta = -2.0295$. Your software will have an option of giving **starting values** to the iterative process. If you give it a starting value of -2, it is highly likely you will end up with the estimate $b = -2.0295$. This value is not the nonlinear least squares estimate, however. The nonlinear least squares estimate is at the global minimum which is the smallest of the minima if more than one exists. How do you guard against ending up at a local minimum? It is wise to try different starting values to ensure you end up at the same place each time. Notice that the curvature at the local minimum in Figure 6.1 is much less than at the global minimum. This should be reflected in a larger "standard error" at the local minimum. Such is indeed the case. We find the outer-product-gradient method yields $\text{se}(b) = 0.3024$, and from the observed-Hessian method we obtain $\text{se}(b) = 0.3577$.

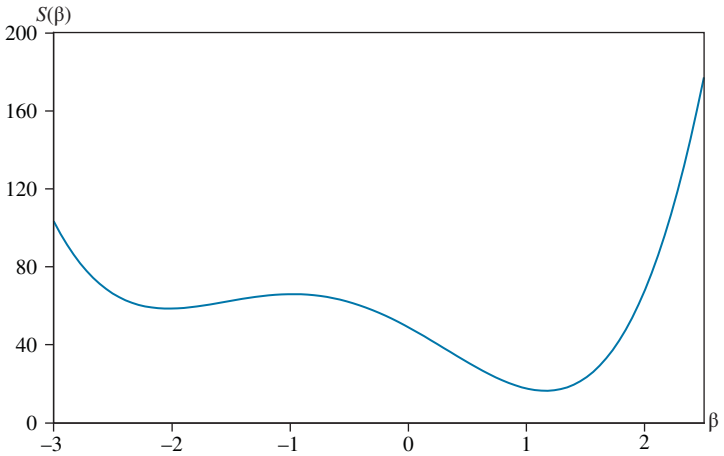


FIGURE 6.1 **Sum of squared errors function for single-parameter example.**

[11] Details of how the numerical algorithms work, how standard errors are obtained, the asymptotic properties of the estimators, and the assumptions necessary for the asymptotic properties to hold, can be found in William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Chapter 7.

[12] These methods require advanced material. See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Section 14.4.6.

EXAMPLE 6.20 | A Logistic Growth Curve

A model that is popular for modeling the diffusion of technological change is the logistic growth curve[13]

$$y_t = \frac{\alpha}{1 + \exp(-\beta - \delta t)} + e_t \tag{6.52}$$

where y_t is the adoption proportion of a new technology. For example, y_t might be the proportion of households who own a computer, or the proportion of computer-owning households who have the latest computer, or the proportion of musical recordings sold as compact disks. In the example that follows, y_t is the share of total U.S. crude steel production that is produced by electric arc furnace technology.

Before considering this example, we note some details about the relationship in equation (6.52). There is only one explanatory variable on the right hand side, namely, time, $t = 1, 2, \ldots, T$. Thus, the logistic growth model is designed to capture the rate of adoption of technological change, or, in some examples, the rate of growth of market share. An example of a logistic curve is depicted in Figure 6.2. In this example, the rate of growth increases at first, to a point of inflection that occurs at $t = -\beta/\delta = 20$. Then the rate of growth declines, leveling off to a saturation proportion given by $\alpha = 0.8$. Since $y_0 = \alpha/(1 + \exp(-\beta))$, the parameter β determines how far the share is below saturation level at time zero. The parameter δ controls the speed at which the point of inflection, and the saturation level, are reached. The curve is such that the share at the point of inflection is $\alpha/2 = 0.4$, half the saturation level.

Traditional technology for steel making, involving blast and oxygen furnaces and the use of iron ore, is being displaced by newer electric arc furnace technology that utilizes scrap steel. This displacement has implications for the suppliers of raw materials such as iron ore. Thus, prediction of the future electric arc furnace share of steel production is of vital importance to mining companies. The file *steel* contains annual data on the electric arc furnace share of U.S. steel production from 1970 to 2015. Using this data to find nonlinear least squares estimates of a logistic growth curve yields the following estimates (standard errors):

$$\hat{\alpha} = 0.8144 \ (0.0511) \quad \hat{\beta} = -1.3777 \ (0.0564)$$
$$\hat{\delta} = 0.0572 \ (0.0043)$$

Quantities of interest are the inflection point at which the rate of growth of the share starts to decline, $-\beta/\delta$; the saturation proportion α; the share at time zero, $y_0 = \alpha/(1 + \exp(-\beta))$; and prediction of the share for various years in the future. In Exercise 6.21, you are invited to find interval estimates for these quantities.

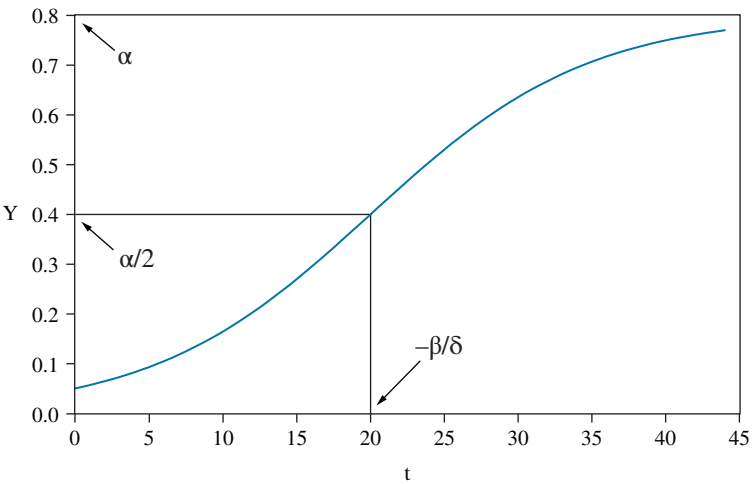


FIGURE 6.2 **A Logistic Growth Curve.**

[13] For other possible models, see Exercises 4.15 and 4.17.

6.7 Exercises

6.7.1 Problems

6.1 When using $N = 50$ observations to estimate the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, you obtain $SSE = 2132.65$ and $s_y = 9.8355$.

a. Find R^2.

b. Find the value of the F-statistic for testing $H_0 : \beta_2 = 0, \beta_3 = 0$. Do you reject or fail to reject H_0 at a 1% level of significance?

c. After augmenting this model with the squares and cubes of predictions $\hat{y}_i^2$ and $\hat{y}_i^3$, you obtain $SSE = 1072.88$. Use RESET to test for misspecification at a 1% level of significance.

d. After estimating the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 z_i^2 + e_i$, you obtain $SSE = 401.179$. What is the R^2 from estimating this model?

e. After augmenting the model in (d) with the squares and cubes of predictions $\hat{y}_i^2$ and $\hat{y}_i^3$, you obtain $SSE = 388.684$. Use RESET to test for misspecification at a 5% level of significance.

6.2 Consider the following model that relates the percentage of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + \beta_5 AGE^2 + e$$

Using 1200 observations from a London survey, this equation was estimated with and without the *AGE* variables included, giving the following results:

$$\underset{\text{(se)}}{\widehat{WALC}} = \underset{}{8.149} + \underset{(0.486)}{2.884 \ln(TOTEXP)} - \underset{(0.382)}{1.217 NK} - \underset{(0.1790)}{0.5699 AGE} + \underset{(0.002332)}{0.005515 AGE^2} \quad \hat{\sigma} = 6.2048$$

$$\underset{\text{(se)}}{\widehat{WALC}} = \underset{}{-1.206} + \underset{(0.482)}{2.152 \ln(TOTEXP)} - \underset{(0.376)}{1.421 NK} \quad \hat{\sigma} = 6.3196$$

a. Use an F-test and a 5% significance level to test whether *AGE* and AGE^2 should be included in the equation.

b. Use an F-test and a 5% significance level to test whether *NK* should be included in the first equation. [*Hint*: $F = t^2$]

c. Use an F-test, a 5% significance level and the first equation to test $H_0 : \beta_2 = 3.5$ against the alternative $H_1 : \beta_2 \neq 3.5$.

d. After estimating the following equation, we find $SSE = 46086$.

$$WALC - 3.5 \ln(TOTEXP) + NK = \beta_1 - (2\beta_5 \times 50) AGE + \beta_5 AGE^2 + e$$

Relative to the original equation with all variables included, for what null hypothesis is this equation the restricted model? Test this null hypothesis at a 5% significance level.

e. What is the χ^2-value for the test in part (d)? In this case, is there a reason why a χ^2-test might be preferred to an F-test?

6.3 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, where $E(e_i|\mathbf{X}) = 0$, $\text{var}(e_i|\mathbf{X}) = \sigma^2$, and $E(e_i e_j|\mathbf{X}) = 0$ for $i \neq j$, with $\mathbf{X}$ representing all observations on x and z. Suppose z_i is omitted from the equation, so that we have the least squares estimator for β_2 as

$$b_2^* = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Prove that

a. $b_2^* = \beta_2 + \beta_3 \sum w_i z_i + \sum w_i e_i$, where $w_i = (x_i - \bar{x}) / \sum (x_i - \bar{x})^2$.

b. $E(b_2^* | \mathbf{X}) = \beta_2 + \beta_3 \widehat{\text{cov}}(x, z) / \widehat{\text{var}}(x)$

c. $\text{var}(b_2^* | \mathbf{X}) = \sigma^2 / (N \widehat{\text{var}}(x))$

d. $\text{var}(b_2^* | \mathbf{X}) \leq \text{var}(b_2 | \mathbf{X})$, where b_2 is the least squares estimator with both x and z included. [*Hint*: check out equation (5.13).]

6.4 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 q_i + e_i$, where $E(e_i | \mathbf{X}) = 0$, with $\mathbf{X}$ representing all observations on x, z, and q. Suppose z_i is unobservable and omitted from the equation, but conditional mean independence $E(z_i | x_i, q_i) = E(z_i | q_i)$ holds, with $E(z_i | q_i) = \delta_1 + \delta_2 q_i$.

a. Show that $E(y_i | x_i, q_i) = (\beta_1 + \beta_3 \delta_1) + \beta_2 x_i + (\beta_3 \delta_2 + \beta_4) q_i$.

b. **i.** Is it possible to get a consistent estimate of the causal effect of x_i on y_i?

ii. Is it possible to get a consistent estimate of the causal effect of z_i on y_i?

iii. Is it possible to get a consistent estimate of the causal effect of q_i on y_i?

6.5 Consider the following wage equation where *EDUC* = years of education and *EXPER* = years of experience:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

Suppose that observations on *EXPER* are not available, and so you decide to use the variables *AGE* and AGE^2 instead. What assumptions are sufficient for the least squares estimate for β_2 to be given a causal interpretation?

6.6 Use an F-test to jointly test the relevance of the two variables *XTRA_X5* and *XTRA_X6* for the family income equation in Example 6.12 and Table 6.1.

6.7 In Example 6.15 a prediction interval for *SALES* from Big Andy's Burger Barn was computed for the settings $PRICE_0 = 6$, $ADVERT_0 = 1.9$. Find point and 95% interval estimates for

$$E(SALES | PRICE = 6, ADVERT = 1.9)$$

Contrast your answers with the point and interval predictions that were obtained in Example 6.15. [*Hint*: The easiest way to calculate the standard error for your point estimate is to utilize some of the calculations given in Example 6.15.]

6.8 Consider the wage equation

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6 (EDUC \times EXPER) + e$$

where the explanatory variables are years of education (*EDUC*) and years of experience (*EXPER*). Estimation results for this equation, and for modified versions of it obtained by dropping some of the variables, are displayed in Table 6.10. These results are from 200 observations in the file *cps5_small*.

a. What restriction on the coefficients of Eqn (A) gives Eqn (B)? Use an F-test to test this restriction. Show how the same result can be obtained using a t-test.

b. What restrictions on the coefficients of Eqn (A) give Eqn (C)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

c. What restrictions on the coefficients of Eqn (B) give Eqn (D)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

d. What restrictions on the coefficients of Eqn (A) give Eqn (E)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

e. Based on your answers to parts (a)–(d), which model would you prefer? Why?

f. Compute the missing AIC value for Eqn (D) and the missing SC value for Eqn (A). Which model is favored by the AIC? Which model is favored by the SC?

TABLE 6.10 Wage Equation Estimates for Exercise 6.8

Variable	Coefficient Estimates and (Standard Errors)				
	Eqn (A)	Eqn (B)	Eqn (C)	Eqn (D)	Eqn (E)
C	0.403	1.483	1.812	2.674	1.256
	(0.771)	(0.495)	(0.494)	(0.109)	(0.191)
$EDUC$	0.175	0.0657	0.0669		0.0997
	(0.091)	(0.0692)	(0.0696)		(0.0117)
$EDUC^2$	−0.0012	0.0012	0.0010		
	(0.0027)	(0.0024)	(0.0024)		
$EXPER$	0.0496	0.0228		0.0314	0.0222
	(0.0172)	(0.0091)		(0.0104)	(0.0090)
$EXPER^2$	−0.00038	−0.00032		−0.00060	−0.00031
	(0.00019)	(0.00019)		(0.00022)	(0.00019)
$EXPER \times EDUC$	−0.001703				
	(0.000935)				
SSE	37.326	37.964	40.700	52.171	38.012
AIC	−1.619	−1.612	−1.562		−1.620
SC		−1.529	−1.513	−1.264	−1.554

6.9 RESET suggests augmenting an existing model with the squares or the squares and higher powers of the predictions $\hat{y}_i$. For example, $\left(\hat{y}_i^2\right)$ or $\left(\hat{y}_i^2, \hat{y}_i^3\right)$ or $\left(\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4\right)$. What would happen if you augmented the model with the predictions $\hat{y}_i$?

6.10 Reconsider Example 6.19 where we used nonlinear least squares to estimate the model $y_i = \beta x_{i1} + \beta^2 x_{i2} + e_i$ by minimizing the sum of squares function $S(\beta) = \sum_{i=1}^{N}\left(y_i - \beta x_{i1} - \beta^2 x_{i2}\right)^2$.

a. Show that $\dfrac{dS}{d\beta} = -2\sum_{i=1}^{N} x_{i1}y_i + 2\beta\left(\sum_{i=1}^{N} x_{i1}^2 - 2\sum_{i=1}^{N} x_{i2}y_i\right) + 6\beta^2\sum_{i=1}^{N} x_{i1}x_{i2} + 4\beta^3\sum_{i=1}^{N} x_{i2}^2$

b. Show that $\dfrac{d^2S}{d\beta^2} = 2\left(\sum_{i=1}^{N} x_{i1}^2 - 2\sum_{i=1}^{N} x_{i2}y_i\right) + 12\beta\sum_{i=1}^{N} x_{i1}x_{i2} + 12\beta^2\sum_{i=1}^{N} x_{i2}^2$

c. Given that $\sum_{i=1}^{N} x_{i1}^2 = 10.422155$, $\sum_{i=1}^{N} x_{i2}^2 = 3.586929$, $\sum_{i=1}^{N} x_{i1}x_{i2} = 4.414097$, $\sum_{i=1}^{N} x_{i1}y_i = 16.528022$, and $\sum_{i=1}^{N} x_{i2}y_i = 10.619469$, evaluate $dS/d\beta$ at both the global minimum $\beta = 1.161207$ and at the local minimum $\beta = -2.029494$. What have you discovered?

d. Evaluate $d^2S/d\beta^2$ at both $\beta = 1.161207$ and $\beta = -2.029494$.

e. At the global minimum, we find $\hat{\sigma}_G = 0.926452$ whereas, if we incorrectly use the local minimum, we find $\hat{\sigma}_L = 1.755044$. Evaluate

$$q = \hat{\sigma}\sqrt{\frac{2}{d^2S/d\beta^2}}$$

at both the global and local minimizing values for β and $\hat{\sigma}$. What is the relevance of these values of q? Go back and check Example 6.19 to see what you have discovered.

6.11 In Example 6.7 we tested the joint null hypothesis

$$H_0: \beta_3 + 3.8\beta_4 = 1,\ \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 80$$

in the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e_i$$

By substituting the restrictions into the model and rearranging variables, show how the model can be written in a form where least squares estimation will yield restricted least squares estimates.

6.12 This exercise uses data on 850 houses sold in Baton Rouge, Louisiana during mid-2005. We will be concerned with the selling price in thousands of dollars (*PRICE*), the size of the house in hundreds of square feet (*SQFT*), the number of bathrooms (*BATHS*), and the number of bedrooms (*BEDS*). Consider the following conditional expectations

$$E(PRICE|BEDS) = \alpha_1 + \alpha_2 BEDS \quad \text{(XR6.12.1)}$$

$$E(PRICE|BEDS, SQFT) = \beta_1 + \beta_2 BEDS + \beta_3 SQFT \quad \text{(XR6.12.2)}$$

$$E(SQFT|BEDS) = \gamma_1 + \gamma_2 BEDS \quad \text{(XR6.12.3)}$$

$$E(PRICE|BEDS, SQFT, BATHS) = \delta_1 + \delta_2 BEDS + \delta_3 SQFT + \delta_4 BATHS \quad \text{(XR6.12.4)}$$

$$E(BATHS|BEDS, SQFT) = \theta_1 + \theta_2 BEDS + \theta_3 SQFT \quad \text{(XR6.12.5)}$$

a. Express α_1 and α_2 in terms of the parameters $(\beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2)$.
b. Express β_1, β_2, and β_3 in terms of the parameters $(\delta_1, \delta_2, \delta_3, \delta_4, \theta_1, \theta_2, \theta_3)$.
c. Use the information in Table 6.11 and a 1% significance level to test whether

$$E(PRICE|BEDS, SQFT, BATHS) = E(PRICE|BEDS)$$

d. Show that the estimates in Table 6.11 satisfy the expressions you derived in parts (a) and (b).
e. Can you explain why the coefficient of *BEDS* changed sign when *SQFT* was added to equation (XR6.12.1).
f. Suppose that $E(BATHS|BEDS) = \lambda_1 + \lambda_2 BEDS$. Use the results in Table 6.11 to find estimates for λ_1 and λ_2.
g. Use the estimates from part (f) and the estimates for equations (XR6.12.3) and (XR6.12.4) to find estimates of α_1 and α_2. Do they agree with the estimates in Table 6.11?
h. Would you view any of the parameter estimates as causal?

TABLE 6.11 **Estimates for House Price Equations for Exercise 6.12**

	Coefficient Estimates and (Standard Errors)				
	(XR6.12.1) *PRICE*	(XR6.12.2) *PRICE*	(XR6.12.3) *SQFT*	(XR6.12.4) *PRICE*	(XR6.12.5) *BATHS*
C	−71.873	−0.1137	−6.7000	−24.0509	0.67186
	(16.502)	(11.4275)	(1.1323)	(11.7975)	(0.06812)
BEDS	70.788	−28.5655	9.2764	−32.649	0.1146
	(5.041)	(4.6504)	(0.3458)	(4.593)	(0.0277)
SQFT		10.7104		9.2648	0.04057
		(0.3396)		(0.4032)	(0.00202)
BATHS				35.628	
				(5.636)	
SSE	8906627	4096699	41930.6	3911896	145.588

6.13 Do gun buybacks save lives? Following the "Port Arthur massacre" in 1996, the Australian government introduced a gun buyback scheme in 1997. The success of that scheme has been investigated by Leigh and Neill.[14] Using a subset of their data on the eight Australian states and territories for the years

[14]Leigh, A. and C. Neill (2010), "Do Gun Buybacks Save Lives? Evidence from Panel Data?", *American Law and Economics Review*, 12(2), p. 509–557.

1980–2006, with 1996 and 1997 omitted, making a total of $N = 200$ observations, we estimate the following model

$$SUIC\_RATE = \beta_1 + \beta_2 GUNRATE + \beta_3 URATE + \beta_4 CITY + \beta_5 YEAR + e$$

Three equations are considered, one where *SUIC_RATE* denotes the firearm suicide rate, one where it represents the non-firearm suicide rate and one for the overall suicide rate, all measured in terms of deaths per million population. For the years after 1997, the variable *GUNRATE* is equal to the number of guns bought back during 1997, per thousand population; it is zero for the earlier years; *URATE* is the unemployment rate, *CITY* is the proportion of the population living in an urban area and *YEAR* is included to capture a possible trend. The estimated equations are given in Table 6.12.

TABLE 6.12 **Estimates for Gun Buyback Equations for Exercise 6.13**

	Coefficient Estimates and (Standard Errors)		
	Firearm Suicide Rate	**Non-firearm Suicide Rate**	**Overall Suicide Rate**
C	1909	–1871	38.37
	(345)	(719)	(779.76)
GUNRATE	–0.223	0.553	0.330
	(0.069)	(0.144)	(0.156)
URATE	–0.485	1.902	1.147
	(0.534)	(1.112)	(1.206)
CITY	–0.628	0.053	–0.576
	(0.057)	(0.118)	(0.128)
YEAR	–0.920	0.976	0.056
	(0.174)	(0.362)	(0.393)
SSE	29745	129122	151890
SSE_R	50641	131097	175562

a. Is there evidence that the gun buyback has reduced firearm suicides? Has there been substitution away from firearms to other means of suicide? Is there a trend in the suicide rate?
b. Is there evidence that greater unemployment increases the suicide rate?
c. Test jointly whether *URATE* and *CITY* contribute to the each of the equations. The sums of squared errors for the equations without these variables are given in the row SSE_R.

6.14 Do gun buybacks save lives? Following the "Port Arthur massacre" in 1996, the Australian government introduced a gun buyback scheme in 1997. As mentioned in Exercise 6.13, the success of that scheme has been investigated by Leigh and Neill. Using a subset of their data on the eight Australian states and territories for the years 1980–2006, with 1996 and 1997 omitted, making a total of $N = 200$ observations, we estimate the following model

$$HOM\_RATE = \beta_1 + \beta_2 GUNRATE + \beta_3 YEAR + e$$

Three equations are considered, one where *HOM_RATE* is the homicide rate from firearms, one where it represent the non-firearm homicide rate and one for the overall homicide rate, all measured in terms of deaths per million population. For the years after 1997, the variable *GUNRATE* is equal to the number of guns bought back during 1997, per thousand population; it is zero for the earlier years; *YEAR* is included to capture a possible trend. The estimated equations are given in Table 6.13.

a. Is there evidence that the gun buyback has reduced firearm homicides? Has there been an increase or a decrease in the homicide rate?
b. Using a joint test on the coefficients of *GUNRATE* and *YEAR*, test whether each of the homicide rates has remained constant over the sample period.

TABLE 6.13 **Estimates for Gun Buyback Equations for Exercise 6.14**

Coefficient Estimates and (Standard Errors)			
	Firearm Homicide Rate	Non-firearm Homicide Rate	Overall Homicide Rate
C	694	1097	1791
	(182)	(816)	(907)
GUNRATE	0.0181	0.0787	0.0968
	(0.0352)	(0.1578)	(0.1754)
YEAR	–0.346	–0.540	–0.886
	(0.092)	(0.410)	(0.456)
SSE	9017	181087	223842
s_y	7.1832	30.3436	34.0273

6.15 The following equation estimates the dependence of *CANS* (the weekly number of cans of brand A tuna sold in thousands) on the price of brand A in dollars (*PRA*) and the prices of two competing brands B and C (*PRB* and *PRC*). The equation was estimated using 52 weekly observations.

$$\widehat{E}(CANS|PRA, PRB, PRC) = 22.96 - 47.08PRA + 9.30PRB + 16.51PRC \qquad SSE = 1358.7$$

a. When *PRB* and *PRC* are omitted from the equation, the sum of squared errors increases to 1513.6. Using a 10% significance level, test whether the prices of the competing brands should be included in the equation. $\left(F_{(0.9,\,2,48)} = 2.417\right)$

b. Consider the following two estimated equations: $\widehat{E}(PRB|PRA) = 0.5403 + 0.3395PRA$ and $\widehat{E}(PRC|PRA) = 0.7708 + 0.0292PRA$. If *PRB* and *PRC* are omitted from the original equation for *CANS*, by how much will the coefficient estimate for *PRA* change? By how much will the intercept estimate change?

c. Find point and 95% interval estimates of $E(CANS|PRA = 0.91, PRB = 0.91, PRC = 0.90)$ using the original equation. The required standard error is 1.58.

d. Find a point estimate for $E(CANS|PRA = 0.91)$ using the equation you constructed in part (b). Can you suggest why the point estimates in (c) and (d) are different? Are there values for *PRB* and *PRC* for which they would be identical?

e. Find a 95% prediction interval for *CANS* when $PRA = 0.91$, $PRB = 0.91$ and $PRC = 0.90$. If you were a statistical consultant to the supermarket selling the tuna, how would you report this interval?

f. When $\widehat{CANS}^2$ is added to the original equation as a regressor the sum of squared errors decreases to 1198.9. Is there any evidence that the equation is misspecified?

6.16 Using 28 annual observations on output (Y), capital (K), labor (L) and intermediate materials (M) for the U.S manufacturing sector, to estimate the Cobb–Douglas production function

$$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \beta_4 \ln(M) + e$$

gave the following results

$$b_2 = 0.1856 \qquad b_3 = 0.3990 \qquad b_4 = 0.4157 \qquad SSE = 0.05699 \qquad s_{\ln(Y)} = 0.23752$$

The standard deviations of the explanatory variables are $s_{\ln(K)} = 0.28108$, $s_{\ln(L)} = 0.17203$, and $s_{\ln(M)} = 0.27505$. The sums of squared errors from running auxiliary regressions on the explanatory variables are (the subscript refers to the dependent variable in the auxiliary regression)

$$SSE_{\ln(K)} = 0.14216 \qquad SSE_{\ln(L)} = 0.02340 \qquad SSE_{\ln(M)} = 0.04199$$

a. Find (i) the standard errors for b_2, b_3, and b_4, (ii) the R^2's for each of the auxiliary regressions, and (iii) the variance inflation factors for b_2, b_3, and b_4.

b. Test the significance of b_2, b_3, and b_4 using a 5% level of significance.

c. Use a 5% level of significance to test the following hypotheses: (i) $H_0: \beta_2 = 0$, $\beta_3 = 0$, (ii) $H_0: \beta_2 = 0, \beta_4 = 0$, (iii) $H_0: \beta_3 = 0, \beta_4 = 0$, and (iv) $H_0: \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$. The restricted sums of squared errors for the first three hypotheses are (i) $SSE_R = 0.0551$, (ii) $SSE_R = 0.08357$ and (iii) $SSE_R = 0.12064$.

d. Comment on the presence and impact of collinearity.

6.7.2 Computer Exercises

6.17 Reconsider Example 6.16 in the text. In that example a number of models were assessed on their within-sample and out-of-sample predictive ability using data in the file *br5*. Of the models considered, the one with the best within-sample performance, as judged by the $\overline{R}^2$, AIC and SC criteria was

$$\ln(PRICE) = \beta_1 + \beta_2 AGE + \beta_3 SQFT + \beta_4 AGE^2 + e \tag{XR6.17}$$

In this exercise we investigate whether we can improve on this function by adding the number of bathrooms (*BATHS*) and the number of bedrooms (*BEDROOMS*). Estimate the equations required to fill in the following table. The models have been numbered from 9 to 12 as extensions of those in Table 6.3. Model 2 is the same as equation (XR6.17). For the subsequent models extra variables are added, with model 12 being the last one considered. For the RMSE values, use the last 100 observations as the hold-out sample. Discuss the results. Include in your discussion a comparison with the results in Table 6.3.

Model	Variables included in addition to those in (XR6.17)	$\overline{R}^2$	AIC	SC	RMSE
2	None				
9	*BATHS*				
10	*BATHS, BEDROOMS*				
11	*BATHS, BEDROOMS* × *SQFT*				
12	*BATHS, BEDROOMS* × *SQFT, BATHS* × *SQFT*				

6.18 Consider Example 6.17 where the rice production function

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA) + \beta_3 \ln(LABOR) + \beta_4 \ln(FERT) + e$$

was estimated using data from the file *rice5*.

a. Using data from 1994 only, contrast the outcomes of the following hypothesis tests.

i. $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$,

ii. $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$,

iii. $H_0: \beta_2 = \beta_3 = 0$ versus $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both β_2 and β_3 are nonzero.

b. Show that the restricted model corresponding to the restriction $\beta_2 + \beta_3 + \beta_4 = 1$ is given by

$$\ln\left(\frac{PROD}{AREA}\right) = \beta_1 + \beta_3 \ln\left(\frac{LABOR}{AREA}\right) + \beta_4 \ln\left(\frac{FERT}{AREA}\right) + e$$

c. Some output from estimating the equation in part (b) using 1994 data is given in Table 6.6. It includes point and interval estimates for β_2, $\text{se}(b_2)$, and a p-value for testing $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$. Describe how these results can be obtained and verify that they are correct.

d. Estimate a constant-returns-to-scale production function using data from both 1993 and 1994. Compare the standard errors and 95% interval estimates with those in Table 6.7 where both years of data were used, but constant returns to scale was not imposed. Include all coefficients in your comparison. What are the auxiliary R^2's for the two variables in the restricted model?

6.19 Consider the following expenditure share equation where *WFOOD* is the proportion of household total expenditure allocated to food, *TOTEXP* is total weekly household expenditure in British pounds (£), and *NK* is the number of children in the household. Conditions MR1–MR5 are assumed to hold. We will be using data from the file *london5*.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 \left[NK \times \ln(TOTEXP)\right] + e$$

a. For a household with the median total expenditure of £90, show that the change in $E(WFOOD|TOTEXP, NK)$ from adding an extra child is $\beta_3 + \beta_4 \ln(90)$.

b. For a household with two children, show that the change in $E(WFOOD|TOTEXP, NK)$ from an increase in total expenditure from £80/week to £120/week is $\beta_2 \ln(1.5) + 2\beta_4 \ln(1.5)$.

c. For a household with two children and total expenditure of £90/week, show that

$$E(WFOOD|TOTEXP, NK) = \beta_1 + \beta_2 \ln(90) + 2\beta_3 + 2\beta_4 \ln(90)$$

d. Consider the following three statements:

A. $\beta_3 + \beta_4 \ln(90) = 0.025$

B. $\beta_2 \ln(1.5) + 2\beta_4 \ln(1.5) = -0.04$

C. $\beta_1 + \beta_2 \ln(90) + 2\beta_3 + 2\beta_4 \ln(90) = 0.37$

We will be concerned with using F and χ^2 tests to test the following three null hypotheses: $H_0^{(1)}$: A is true; $H_0^{(2)}$: A and B are true; $H_0^{(3)}$: A and B and C are true. The alternative hypothesis in each case is that $H_0^{(i)}$ is not true.

What are the relationships between the F and χ^2 tests for each of the three hypotheses? For $H_0^{(1)}$, what is the relationship between the t and F tests?

e. Find the p-values for the F and χ^2 tests for $H_0^{(1)}$, $H_0^{(2)}$, and $H_0^{(3)}$, first using the first 100 observations in *london5*, then using the first 400 observations, and then using all 850 observations.

f. Comment on how changing the sample size, and adding more hypotheses, affects the results of the tests. Are there any dramatic differences between the F-test outcomes and the χ^2-test outcomes?

6.20 In Example 6.18, using 900 observations from the data file *br5*, we identified three potentially influential observations in the estimation of the model

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

Those observations were numbers 150, 411 and 540.

a. Estimate the model with (i) all observations, (ii) observation 150 excluded, (iii) observation 411 excluded, (iv) observation 540 excluded, and (v) observations 150, 411, and 540 excluded. Report the results and comment on their sensitivity to the omission of the observations.

b. Using the estimates from all observations, find the forecast errors corresponding to the within sample predictions at observations 150, 411, and 540.

c. Using the estimates obtained when observation 150 is excluded, find the out-of-sample forecast error for observation 150.

d. Using the estimates obtained when observation 411 is excluded, find the out-of-sample forecast error for observation 411.

e. Using the estimates obtained when observation 540 is excluded, find the out-of-sample forecast error for observation 540.

f. Using the estimates obtained when observations 150, 411, and 540 are excluded, find the out-of-sample forecast errors for observations 150, 411, and 540.

g. Compare the forecast errors obtained in parts (b)–(f) and comment on their sensitivity to the omission of the observations.

6.21 Reconsider Example 6.20 where a logistic growth curve for the share of U.S. steel produced by electric arc furnace (EAF) technology was estimated. The curve is given by the equation

$$y_t = \frac{\alpha}{1 + \exp(-\beta - \delta t)} + e_t$$

a. Find 95% interval estimates for the following:

i. The saturation level α.

ii. The inflection point $t_I = -\beta/\delta$ at which the rate of growth starts to decline. What years does the interval correspond to?

iii. The EAF share in 1969.

iv. The predicted EAF shares from 2016 to 2050. Plot the predictions and their 95% bounds. Comment on how far the predictions are from the saturation level and on the behavior of the 95% bounds.

b. Use a 5% significance level to test the joint null hypothesis that the saturation level is 0.85 and the point of inflection is at $t_I = 25$. Set up the null hypothesis for the point of inflection so that it is linear in the parameters β and δ. Given the interval estimates you found in (a)(i) and (a)(ii), does

the result surprise you? What extra information does the test use that was not used in (a)(i) and (a)(ii)?

c. Estimate the model with the restrictions implied by the null hypothesis in (b) imposed. Find the sum of squared errors and test the null hypothesis with an F-test that uses the restricted and unrestricted sums of squared errors. How does this result compare with that from the automatic test command that you used for part (b)?

6.22 To examine the quantity theory of money, Brumm[15] specifies the equation

$$INFLAT = \beta_1 + \beta_2 MONEY + \beta_3 OUTPUT + e$$

where *INFLAT* is the growth rate of the general price level, *MONEY* is the growth rate of the money supply, and *OUTPUT* is the growth rate of national output. According to theory we should observe that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$. The data in the data file *brumm* are on 76 countries for the year 1995.

a. Using a 5% significance level, test

i. the *strong* joint hypothesis that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$.

ii. the *weak* joint hypothesis $\beta_2 = 1$ and $\beta_3 = -1$.

b. Using the DFFITS criterion, find the four most influential observations.

c. Repeat the two tests with the four most influential observations omitted. Does omission of these four observations change the test outcome?

d. Moroney[16] has argued that β_2 is likely to be different for different countries. Suppose that $\beta_2 = \alpha_1 + \alpha_2 MONEY + \alpha_3 OUTPUT$. Substitute this equation into the original model and, omitting the same four influential observations, estimate the new model.

e. Repeat the two tests for the model estimated in (d) for a hypothetical country with the sample median values $MONEY = 16.35$ and $OUTPUT = 2.7$.

6.23 For two inputs X_1 and X_2 and output Y, a constant elasticity of substitution (CES) production function is given by

$$Y = \alpha\left[\delta X_1^{-\rho} + (1-\delta) X_2^{-\rho}\right]^{-\eta/\rho}$$

where $\alpha > 0$ is an efficiency parameter, $\eta > 0$ is a returns to scale parameter, $\rho > -1$ is a substitution parameter, and $0 < \delta < 1$ is a distribution parameter that relates the share of output to each of the two inputs. The elasticity of substitution between the two inputs is given by $\varepsilon = 1/(1+\rho)$. If $\eta = 1$ and $\rho = 0$ ($\varepsilon = 1$), then the CES production function simplifies to the constant-returns-to-scale Cobb–Douglas production function $Y = \alpha X_1^{\delta} X_2^{1-\delta}$.[17] Using the data in the file *rice5*, define $Y = PROD/AREA$, $X_1 = LABOR/AREA$ and $X_2 = FERT/AREA$.

a. Using nonlinear least squares, estimate the following log form of the CES function

$$\ln(Y) = \beta - \frac{\eta}{\rho}\ln\left[\delta X_1^{-\rho} + (1-\delta) X_2^{-\rho}\right] + e$$

where $\beta = \ln(\alpha)$. Report your results and standard errors. [*Hint*: If you run into difficulties, try using 0.5 as the starting value for all of your parameters.]

b. Find 95% interval estimates for α, η, ε, and δ.

c. Using a 5% significance level, test the null hypothesis $H_0: \eta = 1$, $\rho = 0$ against the alternative $H_1: \eta \neq 1$ or $\rho \neq 0$. Does a constant-returns-to-scale Cobb–Douglas function appear to be adequate?

6.24 Using the data in the file *br5*, find least squares estimates of the following house-price relationships for houses sold in Baton Rouge during 2005.

$$\ln(PRICE) = \alpha_1 + \alpha_2 BEDROOMS + e_1$$
$$\ln(PRICE) = \beta_1 + \beta_2 BEDROOMS + \beta_3 SQFT + e_2$$
$$SQFT = \gamma_1 + \gamma_2 BEDROOMS + u_1$$

[15] Brumm, H.J. (2005) "Money Growth, Output Growth, and Inflation: A Reexamination of the Modern Quantity Theory's Linchpin Prediction," *Southern Economic Journal*, 71(3), 661–667.

[16] Moroney, J.R. (2002), "Money Growth, Output Growth and Inflation: Estimation of a Modern Quantity Theory," *Southern Economic Journal*, 69(2), 398–413.

[17] Proving this result requires some advanced calculus. You need to take natural logarithms of both sides, set $\eta = 1$ and use l'Hôpital's rule to take limits as $\rho \to 0$.

a. Report the coefficient estimates and their standard errors.

b. Show how the estimates $(\hat{\alpha}_1, \hat{\alpha}_2)$ can be found from the parameter estimates in the other two equations. How does the interpretation of $\hat{\beta}_2$ differ from the interpretation of $\hat{\alpha}_2$? What would you characterize as the omitted variable bias when estimating α_2? Is there evidence that *BEDROOMS* has a direct effect on ln(*PRICE*)?

c. Estimate the equation $\ln(PRICE) = \theta_1 + \theta_2 SQFT + e_3$. Compare the estimates $\hat{\theta}_2$ and $\hat{\beta}_3$. What was the effect of omitting *BEDROOMS* on the estimated coefficient for *SQFT*? What assumption about e_3 is necessary for θ_2 to be given the causal interpretation: an increase in house size of 100 square feet leads to a θ_2 increase in ln(*PRICE*), when all other variables are held constant?

d. We will investigate whether this assumption might be violated. Estimate the following equation and report the results

$$\ln(PRICE) = \delta_1 + \delta_2 SQFT + \delta_3 AGE + \delta_4 AGE^2 + e_4$$

e. A comparison of this equation with that in part (c) suggests $e_3 = \delta_3 AGE + \delta_4 AGE^2 + e_4$. Assume $E(e_4|SQFT, AGE) = 0$. We wish to investigate whether $E(e_3|SQFT) = 0$. Show that $E(e_3|SQFT) = 0$ if $\delta_3 = \delta_4 = 0$ **or** if $E(AGE|SQFT) = 0$ and $E(AGE^2|SQFT) = 0$.

f. Test the hypothesis $H_0 : \delta_3 = \delta_4 = 0$ at a 5% significance level.

g. Estimate the equations $AGE = \lambda_1 + \lambda_2 SQFT + u_2$ and $AGE^2 = \phi_1 + \phi_2 SQFT + u_3$. Use a 5% significance level to test the hypotheses $H_0 : \lambda_2 = 0$ and $H_0 : \phi_2 = 0$.

h. What do you conclude about the assumption $E(e_3|SQFT) = 0$?

6.25 Using the data in the file *br5*, estimate the equation

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

where *PRICE* is the selling price in thousands of dollars for houses sold in Baton Rouge, Louisiana, in 2005, *SQFT* is the size of each house in hundreds of square feet and *AGE* is the age of each house in years.

a. Report the coefficient estimates and their standard errors.

b. Graph the estimate of $E[\ln(PRICE)|SQFT = 22, AGE]$ against *AGE*. (In the sample the median and average values for *SQFT* are 21.645 and 22.737, respectively.)

c. In part (b), you will have noticed that the higher-priced houses are the very new ones and the very old ones. Using a 5% significance level test the joint null hypothesis that (i) two houses of the same size, a 5-year old house and an 80-year old house, have the same expected log-price, *and* (ii) a 5-year old house with 2000 square feet has the same expected log-price as a 30-year old house with 2800 square feet.

d. Using a 5% significance level, test the joint null hypothesis that (i) houses start becoming more expensive with age when they are 50 years old, *and* (ii) a 2200 square feet house that is 50 years old has an expected log-price that corresponds to $100,000.

e. Add the variables *BATHS* and $SQFT \times BEDROOMS$ to the model with coefficients β_5 and β_6, respectively. Estimate this model and report the results.

f. Using a 5% significance level, test whether adding these two variables has improved the predictive ability of the model.

g. You are building a new 2300 square-feet house ($AGE = 0$) with three bedrooms and two bathrooms. Adding one extra bedroom and bathroom will increase its size by 260 square feet. Estimate the increase in value of the house from the extra bedroom and bathroom. (Use the natural predictor.)

h. What do you estimate will be the extra value of the house in 20 years' time?

6.26 Each morning between 6:30AM and 8:00AM Bill leaves the Melbourne suburb of Carnegie to drive to work at the University of Melbourne. The time it takes Bill to drive to work (*TIME*), depends on the departure time (*DEPART*), the number of red lights that he encounters (*REDS*), and the number of trains that he has to wait for at the Murrumbeena level crossing (*TRAINS*). Observations on these variables for the 249 working days in 2015 appear in the data file *commute5*. *TIME* is measured in minutes. *DEPART* is the number of minutes after 6:30AM that Bill departs. Consider the equation

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

and suppose assumptions MR1–MR5 hold.

a. Test the following joint hypotheses using a 5% significance level:
  i. The expected delay from a red light is 1.8 minutes *and* the expected delay from a train is 3.2 minutes.
  ii. The expected delay from a red light is 2 minutes *and* the expected delay from a train is 3 minutes.
  iii. The expected delay from a train is 3.5 minutes *and* the delay from a train is double that from a red light.
  iv. The expected delay from a train is 3.5 minutes *and* the delay from a train is double that from a red light *and* leaving at 7:30AM instead of 7:00AM makes the trip 10 minutes longer.

b. Bill suspects that the later he leaves, the more likely he is to encounter a train. Test this hypothesis at a 5% significance level using estimates from the model

$$E(TRAINS|DEPART, REDS) = \alpha_1 + \alpha_2 DEPART + \alpha_3 REDS$$

Is there any evidence of a relationship between the number of trains and the number of red lights?

c. Show that

$$E(TIME|DEPART, REDS) = \left(\beta_1 + \beta_4\alpha_1\right) + \left(\beta_2 + \beta_4\alpha_2\right) DEPART + \left(\beta_3 + \beta_4\alpha_3\right) REDS$$

d. Regress *TIME* on *DEPART* and *REDS* to get estimates for $\delta_1 = \beta_1 + \beta_4\alpha_1$, $\delta_2 = \beta_2 + \beta_4\alpha_2$, and $\delta_3 = \beta_3 + \beta_4\alpha_3$. Using these estimates and those from (a) and (c), show that $\hat{\delta}_1 = b_1 + b_4\hat{\alpha}_1$, $\hat{\delta}_2 = \left(b_2 + b_4\hat{\alpha}_2\right)$, and $\hat{\delta}_3 = b_3 + b_4\hat{\alpha}_3$, where b_k denotes an OLS estimate from the original equation.

e. Interpret b_2 and $\hat{\delta}_2$. Why are they different? How would you characterize any omitted variable bias?

6.27 It has been claimed that an extra year of experience increases wage by 0.8% and that an extra year of education is worth 14 extra years of experience. Doing the calculation, this would mean an extra year of education increases wage by 11.2%. We will investigate this hypothesis using data in the file *cps5_small*. Only those observations for which years of education exceeds 7 will be used. Perform all tests at a 5% level of significance.

a. Estimate the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$ and jointly test the claims about the marginal effects of *EDUC* and *EXPER*.

b. Use RESET to test the adequacy of the model; perform the test with the squares of the predictions *and* the squares and cubes of the predictions.

c. After estimating the model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EDUC^2 + \beta_5 EXPER^2 + \beta_6(EDUC \times EXPER) + e$$

jointly test the claims about the marginal effects of *EDUC* and *EXPER* at the following levels of *EDUC* and *EXPER*:
  i. $EDUC = 10$, $EXPER = 5$
  ii. $EDUC = 14$, $EXPER = 24$
  iii. $EDUC = 18$, $EXPER = 40$

d. Use RESET to test the adequacy of the model; perform the test with the squares of the predictions *and* the squares and cubes of the predictions.

e. How would you respond to the claim about the marginal effects of *EDUC* and *EXPER*?

6.28 Using time-series data on five different countries, Atkinson and Leigh[18] investigate the impact of the marginal tax rate paid by high-income earners on the level of inequality. A subset of their data can be found in the file *inequality*.

a. Using data on Australia, estimate the equation $SHARE = \beta_1 + \beta_2 TAX + e$ where *SHARE* is the percentage income share of the top 1% of incomes, and *TAX* is the median marginal tax rate (as a percentage) paid on wages by the top 1% of income earners. Interpret your estimate for β_2. Would you interpret this as a causal relationship?

[18] Atkinson, A.B. and A. Leigh (2013), "The Distribution of Top Incomes in Five Anglo-Saxon Countries over the Long Run," *Economic Record*, 89, 1–17.

b. It is generally recognized that inequality was high prior to the great depression, then declined during the depression and World War II, increasing again toward the end of the sample period. To capture this effect, estimate the following model with a quadratic trend

$$SHARE = \alpha_1 + \alpha_2 TAX + \alpha_3 YEAR + \alpha_4 YEAR^2 + e$$

where *YEAR* is defined as $1 = 1921, 2 = 1922, \ldots, 80 = 2000$. Interpret the estimate for α_2. Has adding the trend changed the effect of the marginal tax rate? Can the change in this estimate, or lack of it, be explained by the correlations between *TAX* and *YEAR* and *TAX* and $YEAR^2$?

c. In what year do you estimate that expected *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

d. The top marginal tax rate in 1974 was 64%. Test the hypothesis that, in the year 2000, the expected income share of the top 1% would have been 6% if the marginal tax rate had been 64% at that time.

e. Test jointly the hypothesis in (d) *and* that a marginal tax rate of 64% in 1925 would have led to an expected income share of 6% for the top 1% of income earners.

f. Add the growth rate (*GWTH*) to the equation in part (b) and reestimate. Interpret the estimated coefficient for *TAX*.

g. Using the equation estimated in part (f), estimate the year when *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

h. Using the equation estimated in part (f), test the hypothesis that, in the year 2000, the expected income share of the top 1% would have been 6% if the marginal tax rate had been 64% at that time.

i. Using the equation estimated in part (f), test jointly the hypothesis in (h) *and* that a marginal tax rate of 64% in 1925 would have led to an expected income share of 6% for the top 1% of income earners.

j. Has adding the variable *GWTH* led to substantial changes to your estimates and test results? Can the changes, or lack of them, be explained by the correlations between *GWTH* and the other variables in the equation?

6.29 Using time-series data on five different countries, Atkinson and Leigh investigate the impact of the marginal tax rate paid by high-income earners on the level of inequality. A subset of their data can be found in the file *inequality*.

a. Using data on the United States, estimate the equation $\ln(SHARE) = \beta_1 + \beta_2 TAX + e$ where *SHARE* is the percentage income share of the top 1% of incomes, and *TAX* is the median marginal tax rate (as a percentage) paid on wages by the top 1% of income earners. Interpret your estimate for β_2. Would you interpret this as a causal relationship?

b. It is generally recognized that inequality was high prior to the great depression, then declined during the depression and World War II, increasing again toward the end of the sample period. To capture this effect, estimate the following model with a quadratic trend

$$\ln(SHARE) = \alpha_1 + \alpha_2 TAX + \alpha_3 YEAR + \alpha_4 YEAR^2 + e$$

where *YEAR* is defined as $1 = 1921, 2 = 1922, \ldots, 80 = 2000$. Interpret the estimate for α_2. Has adding the trend changed the effect of the marginal tax rate? Can the change in this estimate, or lack of it, be explained by the correlations between *TAX* and *YEAR* and *TAX* and $YEAR^2$?

c. In what year do you estimate that *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

d. The top marginal tax rate in 1974 was 50%. Test the hypothesis that, in the year 2000, the expected log income share of the top 1% would have been ln(12) if the marginal tax rate had been 50% at that time.

e. Test jointly the hypothesis in (d) *and* that a marginal tax rate of 50% in 1925 would have led to an expected log income share of log(12) for the top 1% of income earners.

f. Add the growth rate (*GWTH*) to the equation in part (b) and re-estimate. Has adding this variable *GWTH* led to substantial changes to your estimates and test results? Can the changes, or lack of them, be explained by the correlations between *GWTH* and the other variables in the equation?

g. Using the results from part (f), find point and 95% interval estimates for the marginal tax rate that would be required to reduce the income share of the top 1% to 12% in 2001, assuming $GWTH_{2001} = 3$.

6.30 Consider a translog production function where output is measured as firm sales and there are three inputs: capital, labor, and materials. This function can be written as

$$LSALES = \beta_C + \beta_K K + \beta_L L + \beta_M M + \beta_{KK} K^2 + \beta_{LL} L^2 + \beta_{MM} M^2 + \beta_{KL}(K \times L) + \beta_{KM}(K \times M) + \beta_{LM}(L \times M) + e$$

where *LSALES* is the log of sales, and K, L, and M are the logs of capital, labor and materials, respectively. The translog function is often known as a flexible functional form, intended to approximate a variety of possible functional forms. There are two hypotheses that are likely to be of interest:

$$H_0^{(1)}: \beta_{KK} = 0, \ \beta_{LL} = 0, \ \beta_{MM} = 0, \ \beta_{KL} = 0, \ \beta_{KM} = 0, \ \beta_{LM} = 0$$

(A Cobb–Douglas function is adequate)

$$H_0^{(2)}: \begin{cases} \beta_K + \beta_L + \beta_M = 1 \\ 2\beta_{KK} + \beta_{KL} + \beta_{KM} = 0 \\ \beta_{KL} + 2\beta_{LL} + \beta_{LM} = 0 \\ \beta_{KM} + \beta_{LM} + 2\beta_{MM} = 0 \end{cases} \quad \text{(constant returns to scale)}$$

The data file *chemical_small* contains observations on 1200 firms in China's chemical industry, taken in the year 2006. It is a subset of the data used by Baltagi, Egger, and Kesina[19].

a. Use these data to estimate the translog production function. Are all the coefficient estimates significant at a 5% level of significance?

b. Test $H_0^{(1)}$ at a 5% level of significance.

c. Test $H_0^{(2)}$ at a 5% level of significance. What would be the test outcome if you used a 1% level of significance?

d. Does RESET suggest the translog function is adequate?

e. Estimate the model with the restrictions implied by constant returns to scale $\left(H_0^{(2)}\right)$ imposed. Obtain estimates and standard errors for all 10 coefficients.

f. Compare the estimates and standard errors from parts (a) and (e).

g. Does RESET suggest the restricted model is adequate?

6.31 Everaert and Pozzi[20] develop a model to examine the predictability of consumption growth in 15 OECD countries. Their data is stored in the file *oecd*. The variables used are growth in real per capita private consumption (*CSUMPTN*), growth in real per capita government consumption (*GOV*), growth in per capita hours worked (*HOURS*), growth in per capita real disposable labor income (*INC*), and the real interest rate (*R*). Using only the data for Japan, answer the following questions:

a. Estimate the following model and report the results

$$CSUMPTN = \beta_1 + \beta_2 HOURS + \beta_3 GOV + \beta_4 R + \beta_5 INC + e$$

Are there any coefficient estimates that are not significantly different from zero at a 5% level?

b. The coefficient β_2 could be positive or negative depending on whether hours worked and private consumption are complements or substitutes. Similarly, β_3 could be positive or negative depending on whether government consumption and private consumption are complements or substitutes. What have you discovered? What does a test of the hypothesis $H_0: \beta_2 = 0, \beta_3 = 0$ reveal?

c. Re-estimate the equation with *GOV* omitted and, for the coefficients of the remaining variables, comment on any changes in the estimates and their significance.

d. Estimate the equation

$$GOV = \alpha_1 + \alpha_2 HOURS + \alpha_3 R + \alpha_4 INC + v$$

and use these estimates to reconcile the estimates in part (a) with those in part (c).

[19] Baltagi, B.H., P. H. Egger and M. Kesina (2016), "Firm-level Productivity Spillovers in China's Chemical Industry: A Spatial Hausman-Taylor Approach," *Journal of Applied Econometrics*, 31(1), 214–248.

[20] Everaert, G. and L. Ponzi (2014), "The Predictability of Aggregate Consumption Growth in OECD Countries: A Panel Data Analysis," *Journal of Applied Econometrics*, 29(3), 431–453.

e. Re-estimate the models in parts (a) and (c) with the year 2007 omitted and use each of the estimated models to find point and 95% interval forecasts for consumption growth in 2007.

f. Which of the two models, (a) or (c), produced the more accurate forecast for 2007?

6.32 In their study of the prices of Californian and Washington red wines, Costanigro, Mittelhammer and McCluskey[21] categorize the wines into commercial, semipremium, premium, and ultrapremium. Their data for premium wines are stored in the file *wine1*; those for ultrapremium wines are in the file *wine2*. We will be concerned with the variables *PRICE* (bottle price, CPI adjusted), *SCORE* (score out of 100 given by the Wine Spectator Magazine), *AGE* (years of aging), and *CASES* (number of cases produced in thousands).

a. What signs would you expect on the coefficients $(\beta_2, \beta_3, \beta_4)$ in the following model? Why?

$$\ln(PRICE) = \beta_1 + \beta_2 SCORE + \beta_3 AGE + \beta_4 CASES + e$$

b. Estimate separate equations for premium and ultrapremium wine, and discuss the results. Do the coefficients have the expected signs? If not is there an alternative explanation? Is *SCORE* more important for premium wines or ultrapremium wines? Is *AGE* more important for premium wines or ultrapremium wines?

c. Find point and 95% interval estimates for

i. $E[\ln(PRICE)|SCORE = 90, AGE = 2, CASES = 2]$ for premium wines, and

ii. $E[\ln(PRICE)|SCORE = 93, AGE = 3, CASES = 1]$ for ultrapremium wines.

Do the intervals overlap, or is there a clear price distinction between the two classes?

d. Using the "corrected predictor"—see Section 4.5.3—predict the prices for premium and ultrapremium wines for the settings in parts c(i) and c(ii), respectively.

e. Suppose that you are a wine producer choosing between producing 1000 cases of ultrapremium wine that has to be aged three years and is likely to get a score of 93, and 2000 cases of premium wine that is aged two years and is likely to get a score of 90. Which choice gives the higher expected bottle price? Which choice gives the higher expected revenue? (There are 12 bottles in a case of wine.)

6.33 In this exercise we reconsider the premium wine data in the file *wine1*. Please see Exercise 6.32 and *wine1.def* for details.

a. Estimate the following equation using (i) only cabernet wines, (ii) only pinot wines, and (iii) all other varieties:

$$\ln(PRICE) = \beta_1 + \beta_2 SCORE + \beta_3 AGE + \beta_4 CASES + e$$

Using casual inspection, do you think separate equations are needed for the different varieties?

b. We can develop an F-test to test whether there is statistical evidence to suggest the coefficients in the three equations are different. The unrestricted sum of squared errors for such a test is

$$SSE_U = SSE_{CABERNET} + SSE_{PINOT} + SSE_{OTHER}$$

Compute SSE_U.

c. What is the total number of parameters from the three equations? How many parameters are there when we estimate one equation for all varieties? How many parameter restrictions are there if we restrict corresponding coefficients for all varieties to be equal?

d. Estimate one equation for all varieties. This is the restricted model where corresponding coefficients for the different varieties are assumed to be equal.

e. Using a 5% significance level, test whether there is evidence to suggest there should be different equations for different varieties. What is the null hypothesis for this test? Develop some notation that enables you to state the null hypothesis clearly and precisely.

---

[21] Costanigro, M., R.C. Mittelhammer and J.J.McCluskey (2009), "Estimating Class Specific Parametric Models Under Class Uncertainty: Local Polynomial Regression Clustering in an Hedonic Analysis Of Wine Markets" *Journal of Applied Econometrics*, 24(7), 1117–1135.

Appendix 6A The Statistical Power of *F*-Tests

In Appendix 3B, we explored the factors that lead us to reject a null hypothesis about the slope parameter in a simple regression using a *t*-test. The probability of rejecting a false null hypothesis is positively related to the magnitude of the hypothesis error, and the total variation in the explanatory variable, and inversely related to the size of σ^2, the error variance. These are components of the noncentrality parameter, (3B.2), for the *t*-statistic, (3B.1), when the null hypothesis is false.

Here we show that the factors that lead us to reject a false joint null hypothesis are much the same. Consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ under assumptions SR1–SR6. We will test the joint null hypothesis $H_0 : \beta_1 = c_1, \beta_2 = c_2$ using an *F*-test. In practice the test is carried out using (6.4) in the usual way. To study the power of the *F*-test we will test an equivalent joint null hypothesis $H_0 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}, \quad \beta_2 = c_2$. If the first pair of hypotheses is true then the second pair of hypotheses is true and vice versa. They are completely equivalent. This is not what you would do in practice but this approach will lead us to a form of the *F*-test that is theoretically useful. In the following steps, we will derive the *F*-statistic by combining test statistics for the separate hypotheses $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$ and $H_0^2 : \beta_2 = c_2$. There are quite a few steps, but do not get discouraged. Each step is small and the reward at the end is substantial. Now is a good time to review Appendix 3B on *t*-tests when the null hypothesis is false, Appendix B.3.6, on the chi-square distribution, Appendix B.3.7, on the *t*-distribution, and Appendix B.3.8, on the *F*-distribution.

If we were going to test the first hypothesis, $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$, what test statistic would we use? Most commonly we use a *t*-test for a single hypothesis. For the present, however, assume that we know the error variance σ^2 so that we also know the true variances and covariance of the least squares estimators that are given in equations (2.14)–(2.16). The test statistic is

$$Z_0^1 = \frac{b_1 + b_2\bar{x} - (c_1 + c_2\bar{x})}{\sqrt{\text{var}(b_1 + b_2\bar{x})}} = \frac{\bar{y} - (c_1 + c_2\bar{x})}{\sqrt{\sigma^2/N}} \tag{6A.1}$$

with Z_0^1 denoting the statistic for the null hypothesis H_0^1. We obtained the second equality by taking advantage of the properties of the least squares estimators, recognizing that $b_1 + b_2\bar{x} = \bar{y}$, and $\text{var}(\bar{y}) = \sigma^2/N$, as shown in Appendix C, equation (C.6). If the null hypothesis is true, Z_0^1 has a standard normal distribution, $N(0,1)$. Our objective is to study testing $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$ when it is not true. To accomplish this rewrite Z_0^1 by adding and subtracting $(\beta_1 + \beta_2\bar{x})$ to the numerator in (6A.1) yielding

$$\begin{aligned} Z_0^1 &= \frac{b_1 + b_2\bar{x} - (\beta_1 + \beta_2\bar{x}) + (\beta_1 + \beta_2\bar{x}) - (c_1 + c_2\bar{x})}{\sqrt{\sigma^2/N}} \\ &= \frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}} + \frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}} \\ &= Z_1 + \delta_1 \end{aligned} \tag{6A.2}$$

The first term, Z_1, has a standard normal distribution; it is the test statistic calculated using the true parameter values,

$$Z_1 = \frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}} \sim N(0,1) \tag{6A.3}$$

The second term

$$\delta_1 = \frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}} \tag{6A.4}$$

is the specification error in the hypothesis $H_0^1: \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$. If the null hypothesis is true then $\delta_1 = 0$. If the null hypothesis H_0^1 is not true, then $\delta_1 \neq 0$, and we must account for the fact that δ_1 depends on the sample values, $\mathbf{x}$. In Appendix B.3.6 we define noncentral chi-square random variables. The random variable $Z_0^1|\mathbf{x} = Z_1 + \delta_1 \sim N(\delta_1, 1)$ and $V_0^1|\mathbf{x} = \left(Z_0^1|\mathbf{x}\right)^2 = \left(Z_1 + \delta_1\right)^2 \sim \chi^2_{(1,\delta_1^2)}$ has a noncentral chi-square distribution with one degree of freedom, and noncentrality parameter $\delta = \delta_1^2$. If the null hypothesis is true then $\delta_1 = 0$ and V_0^1 has the chi-square distribution, $V_0^1 \sim \chi^2_{(1,\delta_1^2=0)} = \chi^2_{(1)}$.

The second piece of the puzzle is similar to the first and follows the steps in Appendix 3B. To test $H_0^2: \beta_2 = c_2$, assuming σ^2 is known, use the test statistic

$$Z_0^2 = \frac{b_2 - c_2}{\sqrt{\text{var}(b_2)}} = \frac{b_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \tag{6A.5}$$

If the null hypothesis is true, Z_0^2 has a standard normal distribution, $N(0,1)$. Our objective is to study testing $H_0^2: \beta_2 = c_2$ when it is not true. To accomplish this rewrite Z_0^2 by adding and subtracting β_2 to the numerator, obtaining

$$Z_0^2 = \frac{b_2 - \beta_2 + \beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} + \frac{\beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} = Z_2 + \delta_2 \tag{6A.6}$$

The first term, Z_2, has a standard normal distribution; it is the test statistic calculated using the true parameter value

$$Z_2 = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \sim N(0, 1) \tag{6A.7}$$

The second term

$$\delta_2 = \frac{\beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \tag{6A.8}$$

is the specification error in the hypothesis $H_0^2: \beta_2 = c_2$. If the null hypothesis is true then $\delta_2 = 0$; if the null hypothesis H_0^2 is not true, then $\delta_2 \neq 0$. The random variable $Z_0^2|\mathbf{x} = Z_2 + \delta_2 \sim N(\delta_2, 1)$ and $V_0^2|\mathbf{x} = \left(Z_0^2|\mathbf{x}\right)^2 = \left(Z_2 + \delta_2\right)^2 \sim \chi^2_{(1,\delta_2^2)}$ has a noncentral chi-square distribution with one degree of freedom, and noncentrality parameter $\delta = \delta_2^2$. If the null hypothesis is true, then $\delta_2 = 0$ and V_0^2 has the chi-square distribution, $V_0^2 \sim \chi^2_{(1,\delta_2^2=0)} = \chi^2_{(1)}$.

What is the distribution of $V_1 = V_0^1 + V_0^2 = \left(Z_1 + \delta_1\right)^2 + \left(Z_2 + \delta_2\right)^2$? If Z_1 and Z_2 are statistically independent then $V_1|\mathbf{x} \sim \chi^2_{(2,\delta)}$ with noncentrality parameter $\delta = \delta_1^2 + \delta_2^2$. Because Z_1 and Z_2 are normally distributed random variables, we can prove they are independent by showing that their correlation, or covariance, is zero. Their covariance is

$$\text{cov}(Z_1, Z_2) = E\left\{ \left[Z_1 - E(Z_1)\right]\left[Z_2 - E(Z_2)\right] \right\} = E(Z_1 Z_2)$$

because Z_1 and Z_2 have zero mean, $E(Z_1) = E(Z_2) = 0$. We will show that $E(Z_1 Z_2|\mathbf{x}) = 0$ from which it follows that $E(Z_1 Z_2) = 0$.

$$E(Z_1 Z_2|\mathbf{x}) = E\left\{\left[\frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}}\right]\left[\frac{b_2 - \beta_2}{\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}}\right]\middle|\mathbf{x}\right\}$$

$$= E\left\{\frac{\sqrt{N}}{\sigma}\left[(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}\right]\frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sigma}(b_2 - \beta_2)\middle|\mathbf{x}\right\} \tag{6A.9}$$

$$= \frac{\sqrt{N}\sqrt{\sum(x_i - \bar{x})^2}}{\sigma^2}E\left\{\left[(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}\right](b_2 - \beta_2)|\mathbf{x}\right\}$$

The key component in the last equality is, using (2.15) and (2.16),

$$E\left[(b_1 - \beta_1)(b_2 - \beta_2) + (b_2 - \beta_2)^2\bar{x}\middle|\mathbf{x}\right] = \left[\text{cov}(b_1, b_2|\mathbf{x}) + \bar{x}\text{var}(b_2|\mathbf{x})\right]$$

$$= \frac{-\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2} + \frac{\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2} = 0$$

Since the covariance between Z_1 and Z_2 is zero, they are statistically independent. Thus, $V_1|\mathbf{x} \sim \chi^2_{(2,\delta)}$ where $\delta = \delta_1^2 + \delta_2^2$ and

$$\delta = \delta_1^2 + \delta_2^2 = \left[\frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}}\right]^2 + \left[\frac{\beta_2 - c_2}{\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}}\right]^2$$

$$= N\left\{\frac{\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2}{\sigma^2}\right\} + \frac{(\beta_2 - c_2)^2\sum_{i=1}^N(x_i - \bar{x})^2}{\sigma^2} \tag{6A.10}$$

The final step is to use V_2 from Section 6.1.5, and that V_1 and V_2 are statistically independent. Following a similar procedure to that in (6.13), we form the F-ratio

$$F|\mathbf{x} = \frac{V_1/2}{V_2/(N-2)} \sim F_{(2,N-2,\delta)}$$

In Figure B.9b we show that increases in the noncentrality parameter δ shifts the F-density to the right, increasing the probability that it exceeds the appropriate critical value F_c, and increasing the probability of rejecting a false null hypothesis.

Examining the noncentrality parameter δ in (6A.10) we first note that $\delta \geq 0$, and $\delta = 0$ only if the joint null hypothesis $H_0 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$, $\beta_2 = c_2$, or $H_0 : \beta_1 = c_1$, $\beta_2 = c_2$, is true. The factors that cause δ to increase are as follows:

1. The magnitude of the hypothesis error. In this example the hypothesis error includes two components, $\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2$ and $(\beta_2 - c_2)^2$. The larger these specification errors the higher the probability that the null hypothesis will be rejected. The first term is related to the intercept parameter where the errors in hypotheses about both β_1 and β_2 are contributors, as well as the sample mean, $\bar{x}$. If the sample mean $\bar{x} = 0$, then only the magnitude of $(\beta_1 - c_1)^2$ matters.

2. The sample size, N. As the sample size N increases the value of δ increases not only because it multiplies the first component of δ but also because the data variation $\sum_{i=1}^{N}(x_i - \bar{x})^2$ increases, or at worst stays the same, as N increases. This is very reassuring and a reason to prefer larger samples to smaller ones. The probability of rejecting a false hypothesis approaches one as $N \to \infty$.
3. The variation in the explanatory variable. In the simple regression model the data variation $\sum_{i=1}^{N}(x_i - \bar{x})^2$ is directly related to the probability of rejecting the joint null hypothesis. The larger the data variation, the smaller the variance of b_2, and the more likely we are to detect the discrepancy between β_2 and the hypothesized value c_2.
4. The error variance σ^2. The smaller the error variance, the smaller the uncertainty in the model, and the larger δ becomes, and the higher the probability of rejecting a false joint hypothesis.

For a numerical example we use values arising from the simulation experiment used in Appendix 2H and Appendix 3B. In the first Monte Carlo sample, data file *mc1_fixed_x*, the x-values consist of $x_i = 10$, $i = 1, \ldots, 20$ and $x_i = 20$, $i = 21, \ldots, 40$. The sample mean is $\bar{x} = 15$ so that $\sum(x_i - \bar{x})^2 = 40 \times 5^2 = 1000$. Also, $\sigma^2 = 2500$. The true parameter values in the simulation experiment are $\beta_1 = 100$ and $\beta_2 = 10$. We now test the joint hypothesis $H_0: \beta_1 = 100, \beta_2 = 9$ against the alternative $H_1: \beta_1 \neq 100$ and/or $\beta_2 \neq 9$. At the 5% level of significance we reject the joint null hypothesis if the F-test statistic is greater than the critical value $F_{(0.95,\,2,\,38)} = 3.24482$. You can confirm that the calculated value of the F-statistic is 4.96, so that, at the 5% level of significance, we correctly reject $H_0: \beta_1 = 100, \beta_2 = 9$.

The noncentrality parameter is

$$\delta = N\left\{\frac{\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2}{\sigma^2}\right\} + \frac{(\beta_2 - c_2)^2\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2}$$

$$= 40\left\{\frac{[(100 - 100) + (10 - 9)\,15]^2}{2500}\right\} + \frac{(10 - 9)^2 1000}{2500} = \frac{(40 \times 15^2) + 1000}{2500}$$

$$= 4$$

The probability of rejecting the joint null hypothesis is the probability that a value from a noncentral F-distribution with noncentrality parameter $\delta = 4$ will exceed $F_{(0.95,\,2,\,38)} = 3.24482$. The test power is $P\left[F_{(m_1=2,\ m_2=38,\ \delta=4)} > 3.24482\right] = 0.38738$.

As another illustration let us test the null hypothesis $H_0: \beta_2 = 9$ against $H_1: \beta_2 \neq 9$ using an F-test. The test critical value is the 95th percentile of the F-distribution, $F_{(0.95,\,1,\,38)} = 4.09817$. The calculated F-test value is 4.91 which exceeds the 5% critical value, so once again we correctly reject the null hypothesis. The noncentrality parameter of the F-distribution for this single hypothesis is the square of δ_2 in (6A.8),

$$\delta = \delta_2^2 = \frac{(\beta_2 - c_2)^2}{\sigma^2/\sum(x_i - \bar{x})^2} = \frac{1}{2500/1000} = 0.4$$

Thus the probability of rejecting the null hypothesis $H_0: \beta_2 = 9$ versus $H_1: \beta_2 \neq 9$ when the true value of $\beta_2 = 10$ is $P\left[F_{(m_1=1,\ m_2=38,\ \delta=0.4)} > 4.09817\right] = 0.09457$.

We note three lessons from this exercise. First, using an F-test, the probability of rejecting the joint hypothesis $H_0: \beta_1 = 100, \beta_2 = 9$ is greater than the probability of rejecting the single hypothesis $H_0: \beta_2 = 9$. Second, in Appendix 3B we found that the probability of rejecting

$H_0: \beta_2 = 9$ versus $H_1: \beta_2 > 9$ using a one-tail t-test was 0.15301, with noncentrality parameter 0.63246. The power of a one-tail test, when it can be appropriately used, is greater than the power of a two-tail test. Third, when using a two-tail t-test the rejection probability must be computed with care because the noncentral t-distribution is not symmetric about zero. The probability of rejecting the hypothesis is

$$P\left(t_{(38,\,0.63246)} \le -1.686\right) + \left[1 - P\left(t_{(38,\,0.63246)} \ge 1.686\right)\right] = 0.0049866 + 0.0895807 = 0.09457$$

Appendix 6B Further Results from the FWL Theorem

In Section 5.2.4, we saw that, from the FWL theorem, the least squares estimate of a coefficient of a particular explanatory variable, say x_2, can be obtained by "partialing out" the effects of the other variables on x_2 and on y, and running a regression with the partialed-out versions of y and x_2. We now consider some further results from the FWL theorem. In particular, we show how the variance of the least squares estimator can be written in terms of a simple expression that depends on x_2 and the partialed-out version of x_2.

Consider the multiple regression model with two explanatory variables, $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$. Partial-out x_3 using the Frisch–Waugh–Lovell (FWL) approach. First, the auxiliary regression of y on x_3 is $y_i = a_1 + a_3 x_{i3} + r_i$ and the least squares residual is $\ddot{y}_i = y_i - \tilde{a}_1 - \tilde{a}_3 x_{i3} = y_i - \tilde{y}_i$, where $\tilde{y}_i = \tilde{a}_1 + \tilde{a}_3 x_{i3}$ is the fitted value from the auxiliary regression. The auxiliary regression of x_2 on x_3 is $x_{i2} = c_1 + c_3 x_{i3} + r_{i2}$ and the least squares residual is $\ddot{x}_{i2} = x_{i2} - \tilde{c}_1 - \tilde{c}_3 x_{i3} = x_{i2} - \tilde{x}_{i2}$, where $\tilde{x}_{i2} = \tilde{c}_1 + \tilde{c}_3 x_{i3}$ is the fitted value from the auxiliary regression for x_2. The FWL theorem says that by estimating the model $\ddot{y}_i = \beta_2 \ddot{x}_{i2} + \ddot{e}_i$, we can obtain the same least squares estimator as from the full model. Because the partialed-out model has no explicit intercept, the least squares estimator is

$$b_2 = \sum \ddot{x}_{i2}\ddot{y}_i \Big/ \sum \ddot{x}_{i2}^2 = \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) \Big/ \sum \left(x_{i2} - \tilde{x}_{i2}\right)^2$$

Note that

- $\tilde{x}_{i2}$ is an estimate of $E\left(x_2 | x_3\right)$ and $\tilde{y}_i$ is an estimate of $E\left(y | x_3\right)$. Thus, when x_3 has been partialed out, we use the conditional means in $b_2 = \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) \Big/ \sum \left(x_{i2} - \tilde{x}_{i2}\right)^2$. When x_3 has not been partialed out we use the unconditional means. A similar statement holds for the variance.
- If we replace $\tilde{y}_i$ by $\bar{y}_i$ and replace $x_{i2} - \tilde{x}_{i2}$ by $x_i - \bar{x}_i$, we have the usual expression for the least squares estimator in the simple regression model.
- Further note that the OLS estimator b_2 in the multiple regression model depends on x_2 and y after removing the linear influence of x_3. In addition, the formula above is valid when the multiple regression model contains any number of variables, with the understanding that $\tilde{y}_i$ and $\tilde{x}_{i2}$ are fitted values from auxiliary regressions containing all explanatory variables except x_2. Very neat!

Let us take the numerator $\sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right)$ and work with it.

$$\begin{aligned} \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) &= \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{a}_1 - \tilde{a}_3 x_{i3}\right) \\ &= \sum \left(x_{i2} - \tilde{x}_{i2}\right) y_i - \tilde{a}_1 \sum \left(x_{i2} - \tilde{x}_{i2}\right) - \tilde{a}_3 \sum \left(x_{i2} - \tilde{x}_{i2}\right) x_{i3} \end{aligned}$$

The term $\sum \left(x_{i2} - \tilde{x}_{i2}\right) = 0$ because it is the sum of least squares residuals from the auxiliary regression that includes an intercept. Also $\sum \left(x_{i2} - \tilde{x}_{i2}\right) x_{i3} = 0$ because least squares residuals are uncorrelated with model explanatory variables. See Exercises 2.1 and 2.3. Thus

$$\sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) = \sum \left(x_{i2} - \tilde{x}_{i2}\right) y_i$$

The resulting simplified estimator b_2 is

$$b_2 = \sum \ddot{x}_{i2}\ddot{y}_i / \sum \ddot{x}_{i2}^2 = \sum (x_{i2} - \tilde{x}_{i2}) y_i / \sum (x_{i2} - \tilde{x}_{i2})^2$$

Computationally this is very nice because it is the estimated least squares coefficient from the model $y_i = \beta_2 \ddot{x}_{i2} + \ddot{e}_i$, where $\ddot{x}_{i2} = x_{i2} - \tilde{x}_{i2}$ is a least squares residual.

Now, as in Chapter 2, we can make theoretical progress by further work on the computational form of the least squares estimator. Substitute $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$ into the computational form and simplify.

$$\begin{aligned} b_2 &= \frac{\sum (x_{i2} - \tilde{x}_{i2}) y_i}{\sum (x_{i2} - \tilde{x}_{i2})^2} = \frac{\sum (x_{i2} - \tilde{x}_{i2})(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i)}{\sum (x_{i2} - \tilde{x}_{i2})^2} \\ &= \frac{1}{\sum (x_{i2} - \tilde{x}_{i2})^2} \left[\sum (x_{i2} - \tilde{x}_{i2})(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i) \right] \\ &= \frac{1}{\sum (x_{i2} - \tilde{x}_{i2})^2} \left[\beta_1 \sum (x_{i2} - \tilde{x}_{i2}) + \beta_2 \sum (x_{i2} - \tilde{x}_{i2}) x_{i2} + \beta_3 \sum (x_{i2} - \tilde{x}_{i2}) x_{i3} + \sum (x_{i2} - \tilde{x}_{i2}) e_i \right] \end{aligned}$$

Again $\sum (x_{i2} - \tilde{x}_{i2}) = 0$ and $\sum (x_{i2} - \tilde{x}_{i2}) x_{i3} = 0$. Now, being clever and using $\sum (x_{i2} - \tilde{x}_{i2}) = 0$, we can say

$$\sum (x_{i2} - \tilde{x}_{i2}) x_{i2} = \sum (x_{i2} - \tilde{x}_{i2}) x_{i2} - \tilde{x}_{i2} \sum (x_{i2} - \tilde{x}_{i2}) = \sum (x_{i2} - \tilde{x}_{i2})^2$$

Plugging all this in, we have

$$b_2 = \beta_2 + \frac{\sum (x_{i2} - \tilde{x}_{i2}) e_i}{\sum (x_{i2} - \tilde{x}_{i2})^2}$$

Then, if errors are homoskedastic and serially uncorrelated

$$\begin{aligned} \operatorname{var}(b_2 | \mathbf{X}) &= \operatorname{var}\left[\frac{\sum (x_{i2} - \tilde{x}_{i2}) e_i}{\sum (x_{i2} - \tilde{x}_{i2})^2} \middle| \mathbf{X} \right] = \frac{\sum (x_{i2} - \tilde{x}_{i2})^2 \operatorname{var}(e_i | \mathbf{X})}{\left[\sum (x_{i2} - \tilde{x}_{i2})^2 \right]^2} = \frac{\sum (x_{i2} - \tilde{x}_{i2})^2 \sigma^2}{\left[\sum (x_{i2} - \tilde{x}_{i2})^2 \right]^2} \\ &= \frac{\sigma^2}{\sum (x_{i2} - \tilde{x}_{i2})^2} \end{aligned}$$

CHAPTER 7

Using Indicator Variables

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to explain

1. The difference between qualitative and quantitative economic variables.
2. How to include a 0–1 indicator variable on the right-hand side of a regression, how this affects model interpretation, and give an example.
3. How to interpret the coefficient on an indicator variable in a log-linear equation.
4. How to include a slope-indicator variable in a regression, how this affects model interpretation, and give an example.
5. How to include a product of two indicator variables in a regression, and how this affects model interpretation, giving an example.
6. How to model qualitative factors with more than two categories (similar to region of the country), and how to interpret the resulting model, giving an example.
7. The consequences of ignoring a structural change in parameters during part of the sample.
8. How to test the equivalence of two regression equations using indicator variables.
9. How to estimate and interpret a regression with an indicator dependent variable.
10. The difference between a randomized controlled experiment and a natural experiment.
11. The difference between the average treatment effect (ATE) and the average treatment effect on the treated (ATT).
12. How to use a regression discontinuity design (RDD), and explain when it is useful.

KEYWORDS

annual indicator variables
average treatment effect
Chow test
dichotomous variables
difference estimator
differences-in-differences estimator
dummy variables
dummy variable trap
exact collinearity
hedonic model
indicator variable
interaction variable
intercept indicator variable
linear probability model
log-linear model
natural experiment
quasi-experiments
reference group
regional indicator variables
regression discontinuity design
seasonal indicator variables
slope-indicator variable
treatment effect

7.1 Indicator Variables

Indicator variables, which were first introduced in Section 2.9, allow us to construct models in which some or all regression model parameters, including the intercept, change for some observations in the sample. To make matters specific, let us consider an example from real estate economics. Buyers and sellers of homes, tax assessors, real estate appraisers, and mortgage bankers are interested in predicting the current market value of a house. A common way to predict the value of a house is to use a **hedonic model**, in which the price of the house is explained as a function of its characteristics, such as its size, location, number of bedrooms, and age. The idea is to break down a good into its component pieces, and then estimate the value of each characteristic.[1]

For the present, let us assume that the size of the house, measured in square feet, *SQFT*, is the only relevant variable in determining house price, *PRICE*. Specify the regression model as

$$PRICE = \beta_1 + \beta_2 SQFT + e \tag{7.1}$$

In this model, β_2 is the value of an additional square foot of living area and β_1 is the value of the land alone.

In real estate, the three most important words are "location, location, and location." How can we take into account the effect of a property's being in a desirable neighborhood, such as one near a university, or near a golf course? Thought of this way, location is a "qualitative" characteristic of a house.

Indicator variables are used to account for qualitative factors in econometric models. They are often called **dummy**, **binary**, or **dichotomous variables** because they take just two values, usually one or zero, to indicate the presence or absence of a characteristic or to indicate whether a condition is true or false. They are also called **dummy variables**, to indicate that we are creating a numeric variable for a qualitative, nonnumeric characteristic. We use the terms *indicator variable* and *dummy variable* interchangeably. Using zero and one for the values of these variables is arbitrary, but very convenient, as we will see. Generally, we define an indicator variable D as

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases} \tag{7.2}$$

Thus, for the house price model, we can define an **indicator variable**, to account for a desirable neighborhood, as

$$D = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases}$$

Indicator variables can be used to capture changes in the model intercept, or slopes, or both. We consider these possibilities in turn.

7.1.1 Intercept Indicator Variables

The most common use of indicator variables is to modify the regression model intercept parameter. Adding the indicator variable D to the regression model, along with a new parameter δ, we obtain

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e \tag{7.3}$$

[1] Such models have been used for many types of goods, including personal computers, automobiles and wine. This famous idea was introduced by Sherwin Rosen (1978) "Hedonic Prices and Implicit Markets," *Journal of Political Economy*, 82, 357–369. The ideas are summarized and applied to asparagus and personal computers in Ernst Berndt (1991) *The Practice of Econometrics: Classic and Contemporary*, Reading, MA: Addison-Wesley, Chapter 4.

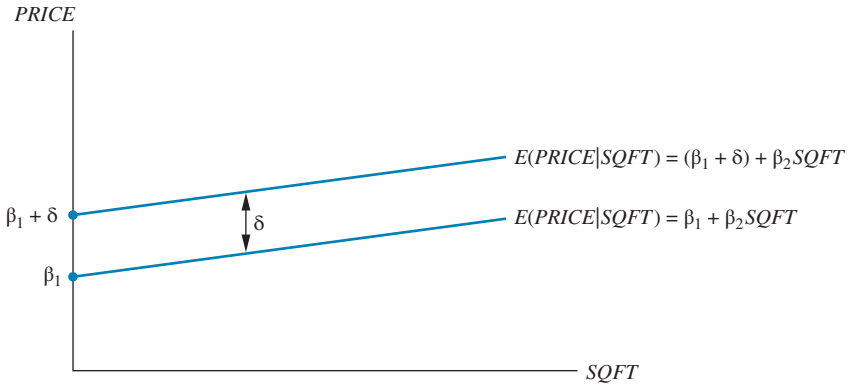


FIGURE 7.1 **An intercept indicator variable.**

The effect of the inclusion of an indicator variable D into the regression model is best seen by examining the regression function, $E(PRICE|SQFT)$, in the two locations. If the model in (7.3) is correctly specified, then $E(e|SQFT, D) = 0$ and

$$E(PRICE|SQFT) = \begin{cases} (\beta_1 + \delta) + \beta_2 SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases} \tag{7.4}$$

In the desirable neighborhood $D = 1$, and the intercept of the regression function is $(\beta_1 + \delta)$. In other areas, the regression function intercept is simply β_1. This difference is depicted in Figure 7.1, assuming that $\delta > 0$.

Adding the indicator variable D to the regression model causes a parallel shift in the relationship by the amount δ. In the context of the house price model the interpretation of the parameter δ is that it is a **location premium**, the difference in house price due to houses being located in the desirable neighborhood. An indicator variable that is incorporated into a regression model to capture a shift in the intercept as the result of some qualitative factor is called an **intercept indicator variable**, or an **intercept dummy variable**. In the house price example, we expect the price to be higher in a desirable location, and thus we anticipate that δ will be positive.

The least squares estimator's properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones—D is treated as any other explanatory variable. We can construct an interval estimate for δ, or we can test the significance of its least squares estimate. Such a test is a statistical test of whether the neighborhood effect on house price is "statistically significant." If $\delta = 0$, then there is no location premium for the neighborhood in question.

Choosing the Reference Group

The convenience of the values $D = 0$ and $D = 1$ is seen in (7.4). The value $D = 0$ defines the **reference group**, or **base group**, of houses that are not in the desirable neighborhood. The expected price of these houses is simply $E(PRICE|SQFT) = \beta_1 + \beta_2 SQFT$. Using (7.3), we are comparing the house prices in the desirable neighborhood to those in the base group.

A researcher can choose whichever neighborhood is most convenient, for expository purposes, to be the reference group. For example, we can define the indicator variable LD to denote the less desirable neighborhood:

$$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$

This indicator variable is defined just the opposite from D, and $LD = 1 - D$. If we include LD in the model specification

$$PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e$$

then we make the reference group, $LD = 0$, the houses in the desirable neighborhood.

You may be tempted to include both D and LD in the regression model to capture the effect of each neighborhood on house prices. That is, you might consider the model

$$PRICE = \beta_1 + \delta D + \lambda LD + \beta_2 SQFT + e$$

In this model, the variables D and LD are such that $D + LD = 1$. Since the intercept variable $x_1 = 1$, we have created a model with **exact collinearity**, and as explained in Section 6.4, the least squares estimator is not defined in such cases. This error is sometimes described as falling into the **dummy variable trap**. By including only one of the indicator variables, either D or LD, the omitted variable defines the reference group, and we avoid the problem.[2]

7.1.2 Slope-Indicator Variables

Instead of assuming that the effect of location on house price causes a change in the intercept of the hedonic regression (7.1), let us assume that the change is in the slope of the relationship. We can allow for a change in a slope by including in the model an additional explanatory variable that is equal to the product of an indicator variable and a continuous variable. In our model, the slope of the relationship is the value of an additional square foot of living area. If we assume that this is one value for homes in the desirable neighborhood, and another value for homes in other neighborhoods, we can specify

$$PRICE = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{7.5}$$

The new variable $(SQFT \times D)$ is the product of house size and the indicator variable, and is called an **interaction variable**, as it captures the interaction effect of location and size on house price. Alternatively, it is called a **slope-indicator variable** or a **slope dummy variable** because it allows for a change in the slope of the relationship. The slope-indicator variable takes a value equal to $SQFT$ for houses in the desirable neighborhood, when $D = 1$, and it is zero for homes in other neighborhoods. Despite its unusual nature, a slope-indicator variable is treated just like any other explanatory variable in a regression model. Examining the regression function for the two different locations best illustrates the effect of the inclusion of the slope-indicator variable into the economic model,

$$E(PRICE|SQFT,\ D) = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) = \begin{cases} \beta_1 + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases}$$

In the desirable neighborhood, the price per additional square foot of a home is $(\beta_2 + \gamma)$; it is β_2 in other locations. We would anticipate $\gamma > 0$ if price per additional square foot is higher in the more desirable neighborhood. This situation is depicted in Figure 7.2a.

Another way to see the effect of including a slope-indicator variable is to use calculus. The partial derivative of expected house price with respect to size (measured in square feet), which gives the slope of the relation, is

$$\frac{\partial E(PRICE|SQFT,\ D)}{\partial SQFT} = \begin{cases} \beta_2 + \gamma & \text{when } D = 1 \\ \beta_2 & \text{when } D = 0 \end{cases}$$

If the assumptions of the regression model hold for (7.5), then the least squares estimators have their usual good properties, as discussed in Section 5.3. A test of the hypothesis that the value of an additional square foot of living area is the same in the two locations is carried out by testing the

[2] Another way to avoid the dummy variable trap is to omit the intercept from the model.

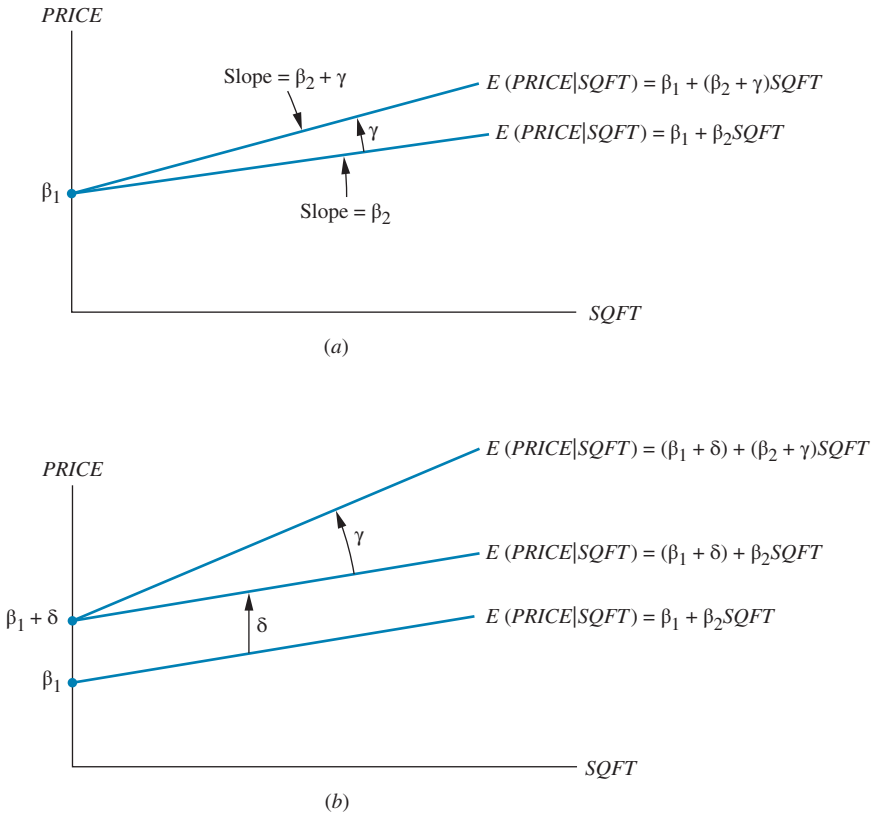


FIGURE 7.2 **(*a*) A slope-indicator variable. (*b*) Slope- and intercept-indicator variables.**

null hypothesis $H_0: \gamma = 0$ against the alternative $H_1: \gamma \neq 0$. In this case, we might test $H_0: \gamma = 0$ against $H_1: \gamma > 0$, since we expect the effect to be positive.

If we assume that house location affects *both* the intercept and the slope, then both effects can be incorporated into a single model. The resulting regression model is

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e \tag{7.6}$$

In this case, the regression functions for the house prices in the two locations are

$$E(PRICE|SQFT) = \begin{cases} (\beta_1 + \delta) + (\beta_2 + \gamma)SQFT & \text{when } D = 1 \\ \beta_1 + \beta_2 SQFT & \text{when } D = 0 \end{cases}$$

In Figure 7.2b, we depict the house price relations assuming that $\delta > 0$ and $\gamma > 0$.

EXAMPLE 7.1 | The University Effect on House Prices

A real estate economist collects information on 1000 house price sales from two similar neighborhoods, one called "University Town" bordering a large state university, and another a neighborhood about three miles from the university. A few of the observations are shown in Table 7.1. The complete data file is *utown*.

House prices are given in \$1000; size (*SQFT*) is the number of hundreds of square feet of living area. For example, the first house sold for \$205,452 and has 2346 square feet of living area. Also recorded are the house *AGE* (in years), location (*UTOWN* = 1 for homes near the university, 0 otherwise), whether the house has a pool (*POOL* = 1 if a pool is present, 0 otherwise) and whether the house has a fireplace (*FPLACE* = 1 if a fireplace is present, 0 otherwise).

TABLE 7.1 Representative Real Estate Data Values

PRICE	*SQFT*	*AGE*	*UTOWN*	*POOL*	*FPLACE*
205.452	23.46	6	0	0	1
185.328	20.03	5	0	0	1
248.422	27.77	6	0	0	0
287.339	23.67	28	1	1	0
255.325	21.30	0	1	1	1
301.037	29.87	6	1	0	1

TABLE 7.2 House Price Equation Estimates

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	24.5000	6.1917	3.9569	0.0001
UTOWN	27.4530	8.4226	3.2594	0.0012
SQFT	7.6122	0.2452	31.0478	0.0000
SQFT × *UTOWN*	1.2994	0.3320	3.9133	0.0001
AGE	−0.1901	0.0512	−3.7123	0.0002
POOL	4.3772	1.1967	3.6577	0.0003
FPLACE	1.6492	0.9720	1.6968	0.0901
$R^2 = 0.8706$	$SSE = 230184.4$			

The economist specifies the regression equation as

$$PRICE = \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \gamma(SQFT \times UTOWN) + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e \qquad (7.7)$$

We anticipate that all the coefficients in this model will be positive except β_3, which is an estimate of the effect of age, or depreciation, on house price. Note that *POOL* and *FPLACE* are intercept dummy variables. By introducing these variables we are asking whether, and by how much, these features change house price. Because these variables stand alone, and are not interacted with *SQFT* or *AGE*, we are assuming that they affect the regression intercept, but not the slope. The estimated regression results are shown in Table 7.2. The goodness-of-fit statistic is $R^2 = 0.8706$, indicating that the model fits the data well. The slope-indicator variable is $SQFT \times UTOWN$. Based on one-tail *t*-tests of significance,[3] at the $\alpha = 0.05$ level we reject zero null hypotheses for each of the parameters and accept the alternatives that they are positive, except for the coefficient on *AGE*, which we accept to be negative. In particular, based on these *t*-tests, we conclude that houses near the university have a significantly higher base price, and that their price per additional square foot is significantly higher than in the comparison neighborhood.

The estimated regression function for the houses near the university is

$$\begin{aligned}\widehat{PRICE} &= (24.5 + 27.453) + (7.6122 + 1.2994)SQFT \\ &\quad - 0.1901 AGE + 4.3772 POOL + 1.6492 FPLACE \\ &= 51.953 + 8.9116 SQFT - 0.1901 AGE \\ &\quad + 4.3772 POOL + 1.6492 FPLACE\end{aligned}$$

For houses in other areas, the estimated regression function is

$$\begin{aligned}\widehat{PRICE} &= 24.5 + 7.6122 SQFT - 0.1901 AGE \\ &\quad + 4.3772 POOL + 1.6492 FPLACE\end{aligned}$$

Based on the regression results in Table 7.2, we estimate that

- The location premium for lots near the university is \$27,453.
- The change in expected price per additional square foot is \$89.12 for houses near the university and \$76.12 for houses in other areas.
- Houses depreciate \$190.10 per year.
- A pool increases the value of a home by \$4,377.20.
- A fireplace increases the value of a home by \$1,649.20.

[3] Recall that the *p*-value for a one-tail test is half of the reported two-tail *p*-value, providing that the coefficient estimate has the "correct" sign.

7.2 Applying Indicator Variables

Indicator variables can be used to ask and answer a rich variety of questions. In this section, we consider some common applications.

7.2.1 Interactions Between Qualitative Factors

We have seen how indicator variables can be used to represent qualitative factors in a regression model. Intercept indicator variables for qualitative factors are *additive*. That is, the effect of each qualitative factor is added to the regression intercept, and the effect of any indicator variable is independent of any other qualitative factor. Sometimes, however, we might question whether the effects of qualitative factors are independent.

For example, suppose we are estimating a wage equation, in which an individual's wages are explained as a function of their experience, skill, and other factors related to productivity. It is customary to include indicator variables for race and sex in such equations. If we have modeled productivity attributes well, and if wage determination is not discriminatory, then the coefficients of the race and sex indicator variables should not be significant. Including just race and sex indicator variables, however, will not capture interactions between these qualitative factors. Is there a differential in wages for black women? Separate indicator variables for being "black" and "female" will not capture this extra interaction effect. To allow for such a possibility, consider the following specification, in which for simplicity we use only education (*EDUC*) as a productivity measure:

$$\begin{aligned} WAGE = {} & \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE \\ & + \gamma(BLACK \times FEMALE) + e \end{aligned} \tag{7.8}$$

where *BLACK* and *FEMALE* are indicator variables, and thus so is their interaction. These are intercept dummy variables because they are not interacted with any continuous explanatory variable. They have the effect of causing a parallel shift in the regression, as in Figure 7.1. When multiple dummy variables are present, and especially when there are interactions between indicator variables, it is important for proper interpretation to write out the regression function, $E(WAGE|EDUC)$, for each indicator variable combination:

$$E(WAGE|EDUC) = \begin{cases} \beta_1 + \beta_2 EDUC & WHITE - MALE \\ (\beta_1 + \delta_1) + \beta_2 EDUC & BLACK - MALE \\ (\beta_1 + \delta_2) + \beta_2 EDUC & WHITE - FEMALE \\ (\beta_1 + \delta_1 + \delta_2 + \gamma) + \beta_2 EDUC & BLACK - FEMALE \end{cases}$$

In this specification, white males are the reference group because this is the group defined when all indicator variables take the value zero, in this case $BLACK = 0$ and $FEMALE = 0$. The parameter δ_1 measures the effect of being black, relative to the reference group; the parameter δ_2 measures the effect of being female, and the parameter γ measures the effect of being black and female.

EXAMPLE 7.2 | The Effects of Race and Sex on Wage

Using CPS data (data file *cps5_small*) from 2013, we obtain the results in Table 7.3. Holding the effect of education constant, we estimate that on average black males earn \$2.07 per hour less than white males, white females earn \$4.22 less than white males, and black females earn \$5.76 less than white males. The coefficients of *EDUC* and *FEMALE* are significantly different from zero using individual *t*-tests. The coefficient of *BLACK* and the interaction effect between

BLACK and *FEMALE* are not estimated very precisely using this sample of 1200 observations, and are not statistically significant.[4]

Suppose we are asked to test the joint significance of all the qualitative factors. How do we test the hypothesis that neither a person's race nor sex affects wages? We do it by testing the joint null hypothesis $H_0: \delta_1 = 0, \delta_2 = 0, \gamma = 0$ against the alternative that at least one of the tested parameters is not zero. If the null hypothesis is true, race and sex fall out of the regression, and thus have no effect on wages.

To test this hypothesis, we use the F-test procedure that is described in Section 6.1. The test statistic for a joint hypothesis is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

where SSE_R is the sum of squared least squares residuals from the "restricted" model in which the null hypothesis is assumed to be true, SSE_U is the sum of squared residuals from the original, "unrestricted," model, J is the number of joint hypotheses, and $N - K$ is the number of degrees of freedom in the unrestricted model. If the null hypothesis is true, then the test statistic F has an F-distribution with J numerator degrees of freedom and $N - K$ denominator degrees of freedom, $F_{(J, N-K)}$. We reject the null hypothesis if $F \geq F_c$, where F_c is the critical value, illustrated in Figure B.9, for the level of significance α. To test the $J = 3$ joint null hypotheses $H_0: \delta_1 = 0, \delta_2 = 0, \gamma = 0$, we obtain the unrestricted sum of squared errors $SSE_U = 214400.9$ from the model reported in Table 7.3. The restricted sum of squares is obtained by estimating the model that assumes the null hypothesis is true, leading to the fitted model

$$\widehat{WAGE} = -10.4000 + 2.3968EDUC$$
$$\text{(se)} \qquad (1.9624) \quad (0.1354)$$

which has $SSE_R = 220062.3$. The degrees of freedom $(N - K) = (1200 - 5) = 1195$ come from the unrestricted model. The value of the F-statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(220062.3 - 214400.9)/3}{214400.9/1195} = 10.52$$

The 1% critical value [i.e., the 99th percentile value] is $F_{(0.99,\,3,\,1195)} = 3.798$. Thus, we conclude that a worker's race and/or sex affect the wage equation.

TABLE 7.3 **Wage Equation with Race and Sex**

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−9.4821	1.9580	−4.8428	0.0000
EDUC	2.4737	0.1351	18.3096	0.0000
BLACK	−2.0653	2.1616	−0.9554	0.3396
FEMALE	−4.2235	0.8249	−5.1198	0.0000
BLACK × *FEMALE*	0.5329	2.8020	0.1902	0.8492
$R^2 = 0.2277$	$SSE = 214400.9$			

7.2.2 Qualitative Factors with Several Categories

Many qualitative factors have more than two categories. An example is the variable region of the country in our wage equation. The CPS data record worker residence within one of the four regions: northeast, midwest, south, and west. Again, using just the simple wage specification for illustration, we can incorporate indicator variables into the wage equation as

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e \tag{7.9}$$

[4] Estimating this model using the larger data set *cps5*, which contains 9799 observations, yields a coefficient estimate for *BLACK* of −4.3488 with a t-value of −5.81. Similarly, the coefficient of the interaction variable is 3.0873 with a $t = 3.01$. Both of these are statistically significant. Recall from Sections 2.4 and 5.3 that larger sample sizes lead to smaller standard errors and thus more precise estimation. Labor economists tend to use large data sets so that complex effects and interactions can be estimated precisely. We use the smaller data set as a text example so that results can be replicated with student versions of software.

Notice that we have not included the indicator variables for all regions. Doing so would have created a model in which exact collinearity exists. Since the regional categories are exhaustive, the sum of the **regional indicator variables** is $NORTHEAST + SOUTH + MIDWEST + WEST = 1$. Thus, the "intercept variable" $x_1 = 1$ is an exact linear combination of the region indicators. Recall, from Section 6.4, that the least squares estimator is not defined in such cases. Failure to omit one indicator variable will lead to your computer software returning a message saying that least squares estimation fails. This error is the **dummy variable trap** that we mentioned in Section 7.1.1.

The usual solution to this problem is to omit one indicator variable, which defines a **reference group**, as we shall see by examining the regression function,

$$E(WAGE|EDUC) = \begin{cases} (\beta_1 + \delta_3) + \beta_2 EDUC & WEST \\ (\beta_1 + \delta_2) + \beta_2 EDUC & MIDWEST \\ (\beta_1 + \delta_1) + \beta_2 EDUC & SOUTH \\ \beta_1 + \beta_2 EDUC & NORTHEAST \end{cases}$$

The omitted indicator variable, *NORTHEAST*, identifies the reference group for the equation, to which workers in other regions are compared. It is the group that remains when the regional indicator variables *WEST*, *MIDWEST*, and *SOUTH* are set to zero. Mathematically, it does not matter which indicator variable is omitted; the choice can be made that is most convenient for interpretation. The intercept parameter β_1 represents the base wage for a worker with no education who lives in the northeast. The parameter δ_1 measures the expected wage differential between southern workers relative to those in the northeast; δ_2 measures the expected wage differential between midwestern workers and those in the northeast.

EXAMPLE 7.3 | A Wage Equation with Regional Indicators

Using CPS data in data file *cps5_small*, let us take the specification in Table 7.3 and add the regional indicators *SOUTH*, *MIDWEST*, and *WEST*. The results are in Table 7.4. We estimate that workers in the South earn \$1.65 less per hour than workers in the Northeast, and workers in the Midwest earn \$1.94 less than workers in the Northeast, holding other factors constant. These estimates are not significantly different from zero at the 10% level.[5]

TABLE 7.4 **Wage Equation with Regional Indicator Variables**

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−8.3708	2.1540	−3.8862	0.0001
EDUC	2.4670	0.1351	18.2603	0.0000
BLACK	−1.8777	2.1799	−0.8614	0.3892
FEMALE	−4.1861	0.8246	−5.0768	0.0000
BLACK × *FEMALE*	0.6190	2.8008	0.2210	0.8251
SOUTH	−1.6523	1.1557	−1.4297	0.1531
MIDWEST	−1.9392	1.2083	−1.6049	0.1088
WEST	−0.1452	1.2027	−0.1207	0.9039
$R^2 = 0.2308$	$SSE = 213552.1$			

[5]Using the larger CPS data file, *cps5*, the estimated regional coefficients are (*t*-values in parentheses): *SOUTH* −0.9405 (−2.24), *MIDWEST* −2.4299 (−5.58), and *WEST* 0.0088 (0.02).

How would we test the hypothesis that there are no regional differences? This would be a joint test of the null hypothesis that the coefficients of the regional dummies are all zero. In the context of the CPS data, $SSE_U = 213552.1$ for the wage equation in Table 7.4. Under the null hypothesis, the model in Table 7.4 reduces to that in Table 7.3 where $SSE_R = 214400.9$. This yields an F-statistic value of 1.579. The p-value for this test is 0.1926, so we fail to reject the null hypothesis that there are no regional differences in the wage equation intercept, holding other factors constant.[6]

7.2.3 Testing the Equivalence of Two Regressions

In Section 7.1.2, we introduced both intercept and slope-indicator variables into the hedonic equation for house price. The result was given in (7.6)

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e$$

The regression functions for the house prices in the two locations are

$$E(PRICE|SQFT) = \begin{cases} \alpha_1 + \alpha_2 SQFT & D = 1 \\ \beta_1 + \beta_2 SQFT & D = 0 \end{cases}$$

where $\alpha_1 = \beta_1 + \delta$ and $\alpha_2 = \beta_2 + \gamma$. Figure 7.2b shows that by introducing both intercept and slope-indicator variables, we have essentially assumed that the regressions in the two neighborhoods are completely different. We could obtain the estimates for (7.6) by estimating separate regressions for each of the neighborhoods. In this section, we generalize this idea, which leads to the **Chow test**, named after econometrician Gregory Chow. The Chow test is an F-test for the equivalence of two regressions.

By including an intercept indicator variable and an interaction variable for *each* additional variable in an equation, we allow all coefficients to differ based on a qualitative factor. Consider again the wage equation in (7.8)

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma(BLACK \times FEMALE) + e$$

We might ask "Are there differences between the wage regressions for the south and for the rest of the country?" If there are no differences, then the data from the south and other regions can be pooled into one sample, with no allowance made for differing slope or intercept. How can we test this? We can carry out the test by creating intercept and slope-indicator variables for *every* variable in the model, and then jointly testing the significance of the indicator variable coefficients using an F-test. That is, we specify the model

$$\begin{aligned} WAGE = {} & \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma(BLACK \times FEMALE) \\ & + \theta_1 SOUTH + \theta_2(EDUC \times SOUTH) + \theta_3(BLACK \times SOUTH) \\ & + \theta_4(FEMALE \times SOUTH) + \theta_5(BLACK \times FEMALE \times SOUTH) + e \end{aligned} \tag{7.10}$$

In (7.10) we have twice the number of parameters and variables than in (7.8). We have added five new variables, the *SOUTH* intercept indicator variable and interactions between *SOUTH* and the other four variables, and corresponding parameters. Estimating (7.10) is equivalent to estimating (7.8) twice—once for the southern workers and again for workers in the rest of the country.

[6] Using the larger CPS data file, *cps5*, the $F = 14.7594$ which is significant at the 1% level.

To see this, examine the regression functions. Let **X** represent (*EDUC*, *BLACK*, *FEMALE*, *SOUTH*). Then

$$E(WAGE|\mathbf{X}) = \begin{cases} \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE & SOUTH = 0 \\ \quad + \gamma(BLACK \times FEMALE) & \\ (\beta_1 + \theta_1) + (\beta_2 + \theta_2)EDUC + (\delta_1 + \theta_3)BLACK & \\ \quad + (\delta_2 + \theta_4)FEMALE + (\gamma + \theta_5)(BLACK \times FEMALE) & SOUTH = 1 \end{cases}$$

Note that each variable has a separate coefficient for southern and nonsouthern workers.

EXAMPLE 7.4 | Testing the Equivalence of Two Regressions: The Chow Test

In column (1) of Table 7.5, we report the estimates and standard errors for the fully interacted model (7.10), using the full sample. The base model (7.8) is estimated once for workers outside the south [column (2)] and again for southern workers [column (3)]. Note that the coefficient estimates on the nonsouth data in (2) are identical to those using the full sample in (1). The standard errors differ because the estimates of the error variance, σ^2, differ. The coefficient estimates using only southern workers are obtained from the full model by adding the indicator variable interaction coefficients θ_i to the corresponding nonsouth coefficients. For example, the coefficient estimate for *BLACK* in column (3) is obtained as $\left(\hat{\delta}_1 + \hat{\theta}_3\right) = 1.1276 - 4.6204 = -3.4928$. Similarly, the coefficient on *FEMALE* in column (3) is $\left(\hat{\delta}_2 + \hat{\theta}_4\right) = -4.1520 - 0.1886 = -4.3406$.

Furthermore, note that the sum of squared residuals for the full model in column (1), but for a small rounding error, is the sum of the *SSE* from the two separate regressions

$$SSE_{full} = SSE_{nonsouth} + SSE_{south}$$
$$= 125880.0 + 87893.9 = 213773.9$$

Using this indicator variable approach, we can test for a southern regional difference. We estimate (7.10) and test the joint null hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$$

against the alternative that at least one $\theta_i \neq 0$. This is the Chow test. If we reject this null hypothesis, we conclude that there is some difference in the wage equation in the southern

TABLE 7.5 Comparison of Fully Interacted to Separate Models

	(1) Full sample		(2) Nonsouth		(3) South	
Variable	**Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**
C	−9.9991	2.3872	−9.9991	2.2273	−8.4162	3.8709
EDUC	2.5271	0.1642	2.5271	0.1532	2.3557	0.2692
BLACK	1.1276	3.5247	1.1276	3.2885	−3.4928	3.1667
FEMALE	−4.1520	0.9842	−4.1520	0.9182	−4.3406	1.7097
BLACK × *FEMALE*	−4.4540	4.4858	−4.4540	4.1852	3.6655	4.1832
SOUTH	1.5829	4.1821				
EDUC × *SOUTH*	−0.1714	0.2898				
BLACK × *SOUTH*	−4.6204	4.5071				
FEMALE × *SOUTH*	−0.1886	1.8080				
BLACK × *FEMALE* × *SOUTH*	8.1195	5.8217				
SSE	213774.0		125880.0		87893.9	
N	1200		810		390	

region relative to the rest of the country. The test can also be thought of as comparing the estimates in the nonsouth and south in columns (2) and (3) in Table 7.5.

The test ingredients are the unrestricted $SSE_U = 213774.0$ from the full model in Table 7.5 [or the sum of the *SSE*'s from the two separate regressions], the restricted $SSE_R = 214400.9$ comes from Table 7.3. The test statistic for the $J = 5$ hypotheses is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

$$= \frac{(214400.9 - 213774.0)/5}{213774.0/1190} = 0.6980$$

The denominator degrees of freedom come from the unrestricted model, $N - K = 1200 - 10$. The p-value of this test is $p = 0.6250$, and thus we fail to reject the null hypothesis that the wage regression in the South is no different from that in the rest of the country.[7]

Remark

The usual F-test of a joint hypothesis relies on the assumptions MR1–MR6 of the linear regression model. Of particular relevance for testing the equivalence of two regressions is assumption MR3, that the variance of the error term, $\text{var}(e_i|\mathbf{X}) = \sigma^2$, is the same *for all* observations. If we are considering possibly different slopes and intercepts for parts of the data, it might also be true that the error variances are different in the two parts of the data. In such a case, the usual F-test is not valid. Testing for equal variances is covered in Section 8.2, and the question of pooling in this case is covered in Section 8.4. For now, be aware that we are assuming constant error variances in the calculations above.

7.2.4 Controlling for Time

The earlier examples we have given apply indicator variables to cross-sectional data. Indicator variables are also used in regressions using time-series data, as the following examples illustrate.

Seasonal Indicators Summer means outdoor cooking on barbeque grills. What effect might this have on the sales of charcoal briquettes, a popular fuel for grilling? To investigate, let us define a model with dependent variable y_t = the number of 20-pound bags of Royal Oak charcoal sold in week t at a supermarket. Explanatory variables would include the price of Royal Oak, the price of competitive brands (Kingsford and the store brand), the prices of complementary goods (charcoal lighter fluid, pork ribs, and sausages), and advertising (newspaper ads and coupons). While these standard demand factors are all relevant, we may also find strong seasonal effects. All other things being equal, more charcoal is sold in the warm summer months than in other seasons. Thus, we may want to include either monthly indicator variables (e.g., $AUG = 1$ if month is August, $AUG = 0$ otherwise) or **seasonal indicator variables** (in North America, $SUMMER = 1$ if month = June, July, or August; $SUMMER = 0$ otherwise) into the regression. In addition to these seasonal effects, holidays are special occasions for cookouts. In the United States, these are Memorial Day (last Monday in May), Independence Day (July 4), and Labor Day (first Monday in September). Additional sales can be expected in the week before these holidays, meaning that indicator variables for each should be included into the regression.

[7] The p-value of this test using the larger CPS data set, *cps5*, is 0.7753, so that we again fail to reject the null hypothesis.

Year Indicators In the same spirit as seasonal indicator variables, **annual indicator variables** are used to capture year effects not otherwise measured in a model. The real estate model discussed earlier in this chapter provides an example. Real estate data are available continuously, every month, every year. Suppose we have data on house prices for a certain community covering a 10-year period. In addition to house characteristics, such as those employed in (7.7), the overall price level is affected by demand factors in the local economy, such as population change, interest rates, unemployment rate, and income growth. Economists creating "cost-of-living" or "house price" indexes for cities must include a component for housing that takes the pure price effect into account. Understanding the price index is important for tax assessors, who must reassess the market value of homes in order to compute the annual property tax. It is also important to mortgage bankers and other home lenders, who must reevaluate the value of their portfolio of loans with changing local conditions, as well as to homeowners trying to sell their houses, and to potential buyers as they attempt to agree upon a selling price.

The simplest method for capturing these price effects is to include annual indicator variables (e.g., $D99 = 1$ if year $= 1999$; $D99 = 0$ otherwise) into the hedonic regression model. An example can be found in Exercise 7.3.

Regime Effects An economic regime is a set of structural economic conditions that exist for a certain period. The idea is that economic relations may behave one way during one regime, but may behave differently during another. Economic regimes may be associated with political regimes (conservatives in power, liberals in power), unusual economic conditions (oil embargo, recession, hyperinflation), or changes in the legal environment (tax law changes). An investment tax credit[8] was enacted in 1962 in an effort to stimulate additional investment. The law was suspended in 1966, reinstated in 1970, and eliminated in the Tax Reform Act of 1986. Thus, we might create an indicator variable

$$ITC_t = \begin{cases} 1 & \text{if } t = 1962-1965, 1970-1986 \\ 0 & \text{otherwise} \end{cases}$$

A macroeconomic investment equation might be

$$INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$$

If the tax credit was successful, then $\delta > 0$.

7.3 Log-Linear Models

In Section 4.5, we examined the log-linear model in some detail. In this section, we explore the interpretation of indicator variables in **log-linear models**. Some additional detail is provided in Appendix 7A. Let us consider the log-linear model in (7.11). We do not introduce an error term, and we take *EDUC* and *FEMALE* to be given, in order to simplify the exposition.

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \delta FEMALE \tag{7.11}$$

What is the interpretation of the parameter δ? *FEMALE* is an intercept dummy variable, creating a parallel shift of the log-linear relationship when $FEMALE = 1$. That is,

$$\ln(WAGE) = \begin{cases} \beta_1 + \beta_2 EDUC & MALES \ \ (FEMALE = 0) \\ (\beta_1 + \delta) + \beta_2 EDUC & FEMALES \ \ (FEMALE = 1) \end{cases}$$

[8] Intriligator, Bodkin and Hsiao, *Econometric Models, Techniques and Applications*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall, 1996, p. 53.

But what about the fact that the dependent variable is ln(*WAGE*)? Does that have an effect? The answer is yes—and there are two solutions.

7.3.1 A Rough Calculation

First, take the difference between ln(*WAGE*) of females and males:

$$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \delta$$

Recall from Appendix A.1.6 and equation (A.3) that 100 times the log-difference, 100δ, is approximately the percentage difference.

EXAMPLE 7.5 | Indicator Variables in a Log-Linear Model: The Rough Approximation

Using the data file *cps5_small*, the estimated log-linear model (7.11) is

$$\widehat{\ln(WAGE)} = 1.6229 + 0.1024EDUC - 0.1778FEMALE$$
$$\text{(se)} \qquad (0.0692) \quad (0.0048) \qquad (0.0279)$$

Thus, we would estimate that there is a 17.78% differential between male and female wages. This is quick and simple, but there is an approximation error with a difference this large.

7.3.2 An Exact Calculation

We can overcome the approximation error by doing a little algebra. The wage difference is

$$\ln(WAGE)_{FEMALES} - \ln(WAGE)_{MALES} = \ln\left(\frac{WAGE_{FEMALES}}{WAGE_{MALES}}\right) = \delta$$

using the property of logarithms that $\ln(x) - \ln(y) = \ln(x/y)$. These are natural logarithms, and the antilog is the exponential function,

$$\frac{WAGE_{FEMALES}}{WAGE_{MALES}} = e^{\delta}$$

Subtract 1 from each side (in a tricky way) to obtain

$$\frac{WAGE_{FEMALES}}{WAGE_{MALES}} - \frac{WAGE_{MALES}}{WAGE_{MALES}} = \frac{WAGE_{FEMALES} - WAGE_{MALES}}{WAGE_{MALES}} = e^{\delta} - 1$$

The percentage difference between wages of females and males is $100\left(e^{\delta} - 1\right)\%$. See Appendix 7A for a more detailed approach.

EXAMPLE 7.6 | Indicator Variables in a Log-Linear Model: An Exact Calculation

Using the data *cps5_small*, we estimate the wage differential between males and females to be

$$100\left(e^{\hat{\delta}} - 1\right)\% = 100\left(e^{-0.1778} - 1\right)\% = -16.29\%$$

The approximate standard error for this estimate is 2.34%, which is a calculation that may be provided by your software.

7.4 The Linear Probability Model

Economics is sometimes described as the "theory of choice." Many of the choices we make in life are "either—or" in nature. A few examples include the following:

- A consumer who must choose between Coke and Pepsi
- A married woman who must decide whether to enter the labor market or not
- A bank official must choose to accept a loan application or not
- A high school graduate must decide whether to attend college or not
- A member of Parliament, a Senator, or a Representative must vote for or against a piece of legislation.

To analyze and predict such outcomes using an econometric model, we represent the choice using an indicator variable, the value one if one alternative is chosen and the value zero if the other alternative is chosen. Because we are attempting to explain choice between two alternatives, the indicator variable will be the **dependent** variable rather than an independent variable in a regression model.

To begin, let us represent the variable indicating a choice as

$$y = \begin{cases} 1 & \text{if first alternative is chosen} \\ 0 & \text{if second alternative is chosen} \end{cases}$$

If we observe the choices that a random sample of individuals makes, then y is a random variable. If p is the probability that the first alternative is chosen, then $P[y = 1] = p$. The probability that the second alternative is chosen is $P[y = 0] = 1 - p$. The probability function for the binary indicator variable y is

$$f(y) = p^y(1-p)^{1-y}, \quad y = 0, 1$$

The indicator variable y is said to follow a Bernoulli distribution. The expected value of y is $E(y) = p$, and its variance is $\text{var}(y) = p(1-p)$.

We are interested in identifying factors that might affect the probability p using a linear regression function, or, in this context, a **linear probability model**,

$$E(y|\mathbf{X}) = p = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

Proceeding as usual, we break the observed outcome y into a systematic portion, $E(y|\mathbf{X})$, and an unpredictable random error, e, so that the econometric model is

$$y = E(y|\mathbf{X}) + e = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$$

One difficulty with using this model for choice behavior is that the usual error term assumptions cannot hold. The outcome y only takes two values, implying that the error term e also takes only two values, so that the usual "bell-shaped" curve describing the distribution of errors does not hold. The probability functions for y and e are

y value	e value	Probability
1	$1 - (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$	p
0	$-(\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$	$1 - p$

The variance of the error term e is

$$\text{var}(e|\mathbf{X}) = p(1-p) = (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)(1 - \beta_1 - \beta_2 x_2 - \cdots - \beta_K x_K)$$

This error is not homoskedastic, so the usual formula for the variance of the least squares estimator is incorrect. A second problem associated with the linear probability model is that predicted values, $\widehat{E(y)} = \hat{p}$, can fall outside the (0, 1) interval, meaning that their interpretation as probabilities does not make sense. Despite these weaknesses, the linear probability model has the advantage of simplicity, and it has been found to provide good estimates of the marginal effects of changes in explanatory variables x_k on the choice probability p, as long as p is not too close to zero or one.[9]

EXAMPLE 7.7 | The Linear Probability Model: An Example from Marketing

A shopper is deciding between Coke and Pepsi. Define the variable *COKE*:

$$COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$$

The expected value of this variable is $E(COKE|\mathbf{X}) = p_{COKE}$ = probability that Coke is chosen given some conditioning factors. What factors might enter the choice decision? The relative price of Coke to Pepsi (*PRATIO*) is a potential factor. As the relative price of Coke rises, we should observe a reduced probability of its choice. Other factors influencing the consumer might be the presence of store displays for these products. Let *DISP_COKE* and *DISP_PEPSI* be indicator variables taking the value one if the respective store display is present and zero if it is not. We expect that the presence of a Coke display will increase the probability of a Coke purchase, and the presence of a Pepsi display will decrease the probability of a Coke purchase.

The data file *coke*[10] contains "scanner" data on 1140 individuals who purchased Coke or Pepsi. In this sample, 44.7% of the customers chose Coke. The estimated linear probability model is

$$\underset{(\text{se})}{\hat{p}_{COKE}} = \underset{(0.0655)}{0.8902} - \underset{(0.0613)}{0.4009}PRATIO + \underset{(0.0344)}{0.0772}DISP\_COKE - \underset{(0.0356)}{0.1657}DISP\_PEPSI$$

Assuming for the moment that the standard errors are reliable,[11] all the coefficients are significantly different from zero at the $\alpha = 0.05$ level. Recall that *PRATIO* = 1 if the prices of Coke and Pepsi are equal, and that *PRATIO* = 1.10 would represent a case in which Coke was 10% more expensive than Pepsi. Such an increase is estimated to reduce the probability of purchasing Coke by 0.04. A store display for Coke is estimated to increase the probability of a Coke purchase by 0.077, and a Pepsi display is estimated to reduce the probability of a Coke purchase by 0.166. The concerns about predicted probabilities falling outside (0,1) are well founded in general, but in this example only 16 of the 1140 sample observations resulted in predicted probabilities less than zero, and there were no predicted probabilities greater than one.

7.5 Treatment Effects

Consider the question "Do hospitals make people healthier?" Angrist and Pischke[12] report the results of a National Health Interview Survey that included the question "During the past 12 months, was the respondent a patient in a hospital overnight?" Also asked was "Would you say your health in general is excellent, very good, good, fair or poor?" Using the number 1 for poor health and 5 for excellent health, those *who had not* gone to the hospital had an average health score of 3.93, and those *who had been* to the hospital had an average score of 3.21. That is, individuals who had been to the hospital had poorer health than those who had not.

[9] See Chapter 16 for nonlinear models of choice, called probit and logit, which ensure that predicted probabilities fall between zero and one. These models require the use of more complex estimators and methods of inference.

[10] Obtained from the ERIM public data base, James M. Kilts Center, University of Chicago Booth School of Business. *Scanner data* is information recorded at the point of purchase by an electronic device reading a barcode.

[11] The estimates and standard errors are not terribly dissimilar from those obtained using more advanced options discussed in Chapters 8 and 16.

[12] *Mostly Harmless Econometrics: An Empiricist's Guide*, Princeton, 2009, pp. 12–13.

Books on principles of economics warn in the first chapter[13] about the faulty line of reasoning known as **post hoc**, **ergo propter hoc**, which means that one event preceding another does not necessarily make the first the cause of the second. Going to the hospital does not *cause* the poorer health status. Those who were less healthy *chose* to go to the hospital because of an illness or injury, and at the time of the survey were still less healthy than those who had not gone to the hospital. Another way to say this is embodied in the warning that "**correlation** is not the same as **causation**." We observe that those who had been in a hospital are less healthy, but observing this association does not imply that going to the hospital causes a person to be less healthy. Still another way to describe the problem we face in this example is to say that data exhibit a **selection bias** because some people chose (or **self-selected**) to go to the hospital and the others did not. When membership in the treated group is in part determined by choice, then the sample is *not* a random sample. There are systematic factors, in this case health status, contributing to the composition of the sample.

A second example of selection bias may bring the concept closer to home. Are you reading this great book because you are enrolled in an econometrics class? Is the course required, or not? If your class is an "elective," then you and your classmates are *not a random sample* from the broader student population. It is our experience that students taking econometrics as an elective have an ability level and quantitative preparation that is higher, on average, than a random sample from the university population. We also observe that a higher proportion of undergraduate students who take econometrics enroll in graduate programs in economics or related disciplines. Is this a causal relationship? In part, it certainly is, but also your abilities and future plans for graduate training may have drawn you to econometrics, so that the high success rate of our students is in part attributed to **selection bias**.

Selection bias is also an issue when asking

- "How much does an additional year of education increase the wages of married women?" The difficulty is that we are able to observe a woman's wages only if she chooses to join the labor force, and thus the observed data is not a random sample.
- "How much does participation in a job-training program increase wages?" If participation is voluntary, then we may see a greater proportion of less skilled workers taking advantage of such a program.
- "How much does a dietary supplement contribute to weight loss?" If those taking the supplement are among the severely overweight, then the results we observe may not be "typical."

In each of these cases, selection bias interferes with a straightforward examination of the data, and makes more difficult our efforts to measure a **causal effect**, or **treatment effect**.

In some situations, usually those involving the physical or medical sciences, it is clearer how we might study causal effects. For example, if we wish to measure the effect of a new type of fertilizer on rice production, we can **randomly** assign identical rice fields to be treated with a new fertilizer (the **treatment group**), with the others being treated with an existing product (the **control group**). At the end of the growing period, we compare the production on the two types of fields. The key here is that we perform a **randomized controlled experiment**. By randomly assigning subjects to treatment and control groups, we ensure that the differences we observe will result from the treatment. In medical research, the effectiveness of a new drug is measured by such experiments. Test subjects are randomly assigned to the control group, who receive a placebo drug, and the treatment group, who receive the drug being tested. By random assignment of treatment and control groups, we prevent any selection bias from occurring.

As economists, we would like to have the type of information that arises from randomized controlled experiments to study the consequences of social policy changes, such as changes in

[13] See, for example, Campbell R. McConnell and Stanley L. Brue, *Economics, Twelfth Edition*, McGraw-Hill, 1993, pp. 8–9.

laws, or changes in types and amounts of aid and training we provide the poor. The ability to perform randomized controlled experiments is limited because the subjects are people, and their economic well-being is at stake. However, there are some examples. Before we proceed, we will examine the statistical consequences of selection bias for the measurement of treatment effects.

7.5.1 The Difference Estimator

In order to understand the measurement of treatment effects, consider a simple regression model in which the explanatory variable is a dummy variable, indicating whether a particular individual is in the treatment or control group. Let y be the outcome variable, the measured characteristic the treatment is designed to effect. In the rice production example, y would be the output of rice on a particular rice field. Define the indicator variable d as

$$d_i = \begin{cases} 1 & \text{individual in treatment group} \\ 0 & \text{individual in control group} \end{cases} \tag{7.12}$$

The effect of the treatment on the outcome can be modeled as

$$y_i = \beta_1 + \beta_2 d_i + e_i, \quad i = 1, \ldots, N \tag{7.13}$$

where e_i represents the collection of other factors affecting the outcome. The regression functions for the treatment and control groups are

$$E(y_i) = \begin{cases} \beta_1 + \beta_2 & \text{if in treatment group, } d_i = 1 \\ \beta_1 & \text{if in control group, } d_i = 0 \end{cases}$$

This is the same model we used in Section 2.9 to study the effect of location on house prices. The **treatment effect** that we wish to measure is β_2. The least squares estimator of β_2 is

$$b_2 = \frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2} = \bar{y}_1 - \bar{y}_0 \tag{7.14}$$

where $\bar{y}_1 = \sum_{i=1}^{N_1} y_i / N_1$ is the sample mean of the N_1 observations on y for the treatment group ($d = 1$) and $\bar{y}_0 = \sum_{i=1}^{N_0} y_i / N_0$ is the sample mean of the N_0 observations on y for the control group ($d = 0$). In this treatment/control framework, the estimator b_2 is called the **difference estimator** because it is the difference between the sample means of the treatment and control groups.[14]

7.5.2 Analysis of the Difference Estimator

The statistical properties of the difference estimator can be examined using the same strategy employed in Section 2.4.2. We can rewrite the difference estimator as

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)\left(e_i - \bar{e}\right)}{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2} = \beta_2 + \left(\bar{e}_1 - \bar{e}_0\right)$$

[14] See Appendix 7B for an algebraic derivation.

In the middle equality, the factor added to β_2 has the same form as the difference estimator in (7.14), with e_i replacing y_i—hence the final equality. The difference estimator b_2 equals the true treatment effect β_2 plus the difference between the averages of the unobserved factors affecting the outcomes y for the treatment group $(\bar{e}_1)$ and for the control group $(\bar{e}_0)$. In order for the difference estimator to be unbiased, $E(b_2) = \beta_2$, it must be true that

$$E(\bar{e}_1 - \bar{e}_0) = E(\bar{e}_1) - E(\bar{e}_0) = 0$$

In words, the expected value of all the factors affecting the outcome, other than the treatment, must be **equal** for the treatment and control groups.

If we allow individuals to "self-select" into treatment and control groups, then $E(\bar{e}_1) - E(\bar{e}_0)$ is the selection bias in the estimation of the treatment effect. For example, we observed that those who had not gone to the hospital (control group) had an average health score of 3.93, and those who had been to the hospital (treatment group) had an average health score of 3.21. The estimated effect of the treatment is $(\bar{y}_1 - \bar{y}_0) = 3.21 - 3.93 = -0.72$. The estimator bias in this case arises because the preexisting health conditions for the treated group, captured by $E(\bar{e}_1)$, are poorer than the pre-existing health of the control group, captured by $E(\bar{e}_0)$, so that in this example there is a negative bias in the difference estimator.

We can anticipate that anytime some individuals **select** treatment there will be factors leading to this choice that are systematically different from those leading individuals in the control group to not select treatment, resulting in a selection bias in the difference estimator. How can we eliminate the self-selection bias? The solution is to **randomly** assign individuals to treatment and control groups, so that there are no systematic differences between the groups, except for the treatment itself. With random assignment, and the use of a large number of experiment subjects, we can be sure that $E(\bar{e}_1) = E(\bar{e}_0)$ and $E(b_2) = \beta_2$.

EXAMPLE 7.8 | An Application of Difference Estimation: Project STAR

Medical researchers use white mice to test new drugs because these mice, surprisingly, are genetically similar to humans. Mice that are bred to be identical are randomly assigned to treatment and control groups, making estimation of the treatment effect of a new drug on the mice a relatively straightforward and reproducible process. Medical research on humans is strictly regulated, and volunteers are given incentives to participate, then randomly assigned to treatment and control groups. Randomized controlled experiments in the social sciences are equally attractive from a statistician's point of view but are rare because of the difficulties in organizing and funding them. A notable example of a randomized experiment is Tennessee's Project STAR.[15]

A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded, as was some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star*.

Let us first compare the performance of students in small classes versus regular classes.[16]

The variable *TOTALSCORE* is the combined reading and math achievement scores and *SMALL* = 1 if the student was assigned to a small class, and zero if the student is in a regular class. In Table 7.6a and b are summary statistics for the two types of classes. First, note that on all measures **except** *TOTALSCORE* the variable means reported are very

[15] See https://dataverse.harvard.eduldataset.xhtml?persistentld=hdl: 1902.1/10766 for program description, public use data and extensive literature.

[16] Interestingly there is no significant difference in outcomes comparing a regular class to a regular class with an aide. For this example all observations for students in the third treatment group are dropped.

similar. This is because students and teachers were randomly assigned to the classes, so that there should be no patterns evident. The average value of *TOTALSCORE* in the regular classes is 918.0429 and in small classes it is 931.9419, a difference of 13.899 points. The test scores are higher in the smaller classes. The difference estimator obtain using regression will yield the same estimate, along with significance levels.

TABLE 7.6a Summary Statistics for Regular-Sized Classes

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	918.0429	73.1380	635	1229
SMALL	0.0000	0.0000	0	0
TCHEXPER	9.0683	5.7244	0	24
BOY	0.5132	0.4999	0	1
FREELUNCH	0.4738	0.4994	0	1
WHITE_ASIAN	0.6813	0.4661	0	1
TCHWHITE	0.7980	0.4016	0	1
TCHMASTERS	0.3651	0.4816	0	1
SCHURBAN	0.3012	0.4589	0	1
SCHRURAL	0.4998	0.5001	0	1

$N = 2005$

TABLE 7.6b Summary Statistics for Small Classes

Variable	Mean	Std. Dev.	Min	Max
TOTALSCORE	931.9419	76.3586	747	1253
SMALL	1.0000	0.0000	1	1
TCHEXPER	8.9954	5.7316	0	27
BOY	0.5150	0.4999	0	1
FREELUNCH	0.4718	0.4993	0	1
WHITE_ASIAN	0.6847	0.4648	0	1
TCHWHITE	0.8625	0.3445	0	1
TCHMASTERS	0.3176	0.4657	0	1
SCHURBAN	0.3061	0.4610	0	1
SCHRURAL	0.4626	0.4987	0	1

$N = 1738$

The model of interest is

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + e \tag{7.15}$$

The regression results are in column (1) of Table 7.7. The estimated "treatment effect" of putting kindergarten children into small classes is 13.899 points, the same as the difference in sample means computed above, on their achievement score total; the difference is statistically significant at the 0.01 level.

EXAMPLE 7.9 | The Difference Estimator with Additional Controls

Because of the random assignment of the students to treatment and control groups, there is no selection bias in the estimate of the treatment effect. However, if additional factors might affect the outcome variable, they can be included in the regression specification. For example, it is possible that a teacher's experience leads to greater learning and higher achievement test scores. Adding *TCHEXPER* to the base model, we obtain

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + e \tag{7.16}$$

The least squares estimates of (7.16) are in column (2) of Table 7.7. We estimate that each additional year of teaching experience increases the test score performance by 1.156 points, which is statistically significant at the 0.01 level. This increases our understanding of the effect of small classes. The results show that the effect of small classes is the same as the effect of approximately 12 years of teaching experience.

Note that adding *TCHEXPER* to the regression changed the estimate of the effect of *SMALL* classes very little. This is exactly what we would expect if *TCHEXPER* is uncorrelated with *SMALL*. The simple correlation between *SMALL* and *TCHEXPER* is only −0.0064. Recall that omitting a variable that is uncorrelated with an included variable does not change the estimated coefficient of the included variable. Comparing the models in columns (1) and (2) of Table 7.7, the model in (1) omits the significant variable *TCHEXPER*, but there is little change in the estimate of β_2 introduced by omitting this nearly uncorrelated variable. Furthermore, we can expect, in general, to obtain an estimator with smaller standard errors if we are able to include additional controls. In (7.15), any and all factors other than small class size are included in the error term. By taking some of those factors out of the error term and including them in the regression, the variance of the error term σ^2 is reduced, which reduces estimator variance.

TABLE 7.7 **Project STAR: Kindergarten**

	(1)	(2)	(3)	(4)
C	918.0429***	907.5643***	917.0684***	908.7865***
	(1.6672)	(2.5424)	(1.4948)	(2.5323)
SMALL	13.8990***	13.9833***	15.9978***	16.0656***
	(2.4466)	(2.4373)	(2.2228)	(2.2183)
TCHEXPER		1.1555***		0.9132***
		(0.2123)		(0.2256)
SCHOOL EFFECTS	*No*	*No*	*Yes*	*Yes*
N	3743	3743	3743	3743
adj. R^2	0.008	0.016	0.221	0.225
SSE	20847551	20683680	16028908	15957534

Standard errors in parentheses
Two-tail *p*-values: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

EXAMPLE 7.10 | The Difference Estimator with Fixed Effects

It may be that assignment to treatment groups is related to one or more observable characteristics. That is, treatments are randomly assigned *given* an external factor. Prior to a medical experiment concerning weight loss, participants may fall into the "overweight" category and the "obese" category. Of those in the overweight group 30% are randomly assigned for treatment, and of the obese group 50% are randomly assigned for treatment. Given pretreatment status, the treatment is randomly assigned. If such conditioning factors are omitted and put into the error term in (7.15) or (7.16), then these factors are correlated with the treatment variable and the least squares estimator of the treatment effect is biased and inconsistent. The way to adjust to "conditional" randomization is to include the conditioning factors into the regression.

In the STAR data, another factor that we might consider affecting the outcome is the school itself. The students were randomized *within* schools (conditional randomization), but not *across* schools. Some schools may be located in wealthier school districts that can pay higher salaries, thus attracting better teachers. The students in our sample are enrolled in 79 different schools. One way to account for school effects is to include an indicator variable for each school. That is, we can introduce 78 new indicators:

$$SCHOOL\_j = \begin{cases} 1 & \text{if student is in school } j \\ 0 & \text{otherwise} \end{cases}$$

This is an "intercept" indicator variable, allowing the expected total score to differ for each school. The model including these indicator variables is

$$TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 TCHEXPER_i + \sum_{j=2}^{79} \delta_j SCHOOL\_j_i + e_i \qquad (7.17)$$

The regression function for a student in school *j* is

$$E(TOTALSCORE_i|\mathbf{X}) = \begin{cases} (\beta_1 + \delta_j) + \beta_3 TCHEXPER_i & \text{student in regular class} \\ (\beta_1 + \delta_j + \beta_2) + \beta_3 TCHEXPER_i & \text{student in small class} \end{cases}$$

Here **X** represents the variables *SMALL*, *TCHEXPER*, and all the indicator variables *SCHOOL_j*. The expected score for a student in a regular class for a teacher with no experience is adjusted by the fixed amount δ_j. This **fixed effect** controls for some differences in the schools that are not accounted for by the regression model.

Columns (3) and (4) in Table 7.7 contain the estimated coefficients of interest but not the 78 indicator variable coefficients. The joint *F*-test of the hypothesis that all $\delta_j = 0$ consists of $J = 78$ hypotheses with $N - K = 3662$ degrees of freedom. The *F*-value = 14.118 is significant at the 0.001 level. We conclude that there are statistically significant individual differences among schools. The important coefficients on *SMALL* and *TCHEXPER* change a little. The estimated effect of being in a small class increases to 16.0656

achievement test points in model (4), as compared to 13.9833 points in the corresponding model (2). It appears that some effect of small classes was masked by unincorporated individual school differences. This effect is small, however, as the 95% interval estimate for the coefficient of *SMALL* [11.7165, 20.4148] in model (4) includes 13.9833. Similarly, the estimated effect of teacher experience is slightly different in the models with and without the school fixed effects.

EXAMPLE 7.11 | Linear Probability Model Check of Random Assignment

In Table 7.6a and b, we examined the summary statistics for the data sorted by whether pupils were in a regular class or a small class. Except for *TOTALSCORE*, we did not find much difference in the sample means of the variables examined. Another way to check for random assignment is to regress *SMALL* on these characteristics and check for any significant coefficients, or an overall significant relationship. If there is random assignment, we should not find any significant relationships. Because *SMALL* is an indicator variable, we use the linear probability model discussed in Section 7.4. The estimated linear probability model is

$$\widehat{SMALL} = 0.4665 + 0.0014BOY + 0.0044WHITE\_ASIAN$$
$$(t) \qquad\qquad (0.09) \qquad\quad (0.22)$$
$$- 0.0006TCHEXPER - 0.0009FREELUNCH$$
$$(-0.42) \qquad\qquad (-0.05)$$

First, note that none of the right-hand-side variables are statistically significant. Second, the overall F-statistic for this linear probability model is 0.06 with a $p = 0.99$. There is no evidence that students were assigned to small classes based on any of these criteria. Also, recall that the linear probability model is so named because $E(SMALL|\mathbf{X})$ is the probability of observing $SMALL = 1$ in a random draw from the population. If the coefficients of all the potential explanatory factors are zero, the estimated intercept gives the estimated probability of observing a child in a small class to be 0.4665, with 95% interval estimate [0.4171, 0.5158]. We cannot reject the null hypothesis that the intercept equals 0.5, which is what it should be if students are allocated by a "flip" of a coin. *The importance of this, again, is that by randomly assigning students to small classes we can estimate the "treatment" effect using the simple difference estimator in* (7.15). The ability to isolate the important class size effect is a powerful argument in favor of randomized controlled experiments.

7.5.3 The Differences-in-Differences Estimator

Randomized controlled experiments are somewhat rare in economics because they are expensive and involve human subjects. **Natural experiments**, also called **quasi-experiments**, rely on observing real-world conditions that approximate what would happen in a randomized controlled experiment. Treatment appears *as if* it were randomly assigned. In this section, we consider estimating treatment effects using "before and after" data.

Suppose that we observe two groups before and after a policy change, with the **treatment group** being affected by the policy, and the **control group** being unaffected by the policy. Using such data, we will examine any change that occurs to the control group and compare it to the change in the treatment group.

The analysis is explained by Figure 7.3. The outcome variable y might be an employment rate, a wage rate, a price, or so on. Before the policy change we observe the treatment group value $y = B$, and after the policy is implemented the treatment group value is $y = C$. Using only the data on the treatment group we cannot separate out the portion of the change from $y = B$ to $y = C$ that is due to the policy from the portion that is due to other factors that may affect the outcome. We say that the treatment effect is not "identified."

We can isolate the effect of the treatment by using a control group that is not affected by the policy change. Before the policy change, we observe the control group value $y = A$, and after the policy change, the control group value is $y = E$. In order to estimate the treatment effect using the four pieces of information contained in the points A, B, C, and E, we make the

strong assumption that the two groups experience a **common trend**. In Figure 7.3, the dashed line $\overline{BD}$ represents what we imagine the treatment group growth would have been (the term **counterfactual** from psychology is sometimes used to describe this imagined outcome) in the absence of the policy change. The growth described by the dashed line $\overline{BD}$ is unobservable, and is obtained by assuming that the growth in the treatment group that is unrelated to the policy change is the same as the growth in the control group.

The treatment effect $\delta = \overline{CD}$ is the difference between the treatment and control values of y in the "after" period, after subtracting $\overline{DE}$, which is what the difference between the two groups would have been in the absence of the policy. Using the common growth assumption, the difference $\overline{DE}$ equals the initial difference $\overline{AB}$. Using the four observable points A, B, C, and E depicted in Figure 7.3, estimation of the treatment effect is based on data averages for the two groups in the two periods,

$$\begin{aligned}\hat{\delta} &= (\hat{C} - \hat{E}) - (\hat{B} - \hat{A}) \\ &= (\bar{y}_{Treatment,\ After} - \bar{y}_{Control,\ After}) - (\bar{y}_{Treatment,\ Before} - \bar{y}_{Control,\ Before})\end{aligned} \tag{7.18}$$

In (7.18), the sample means are

$$\begin{aligned}\bar{y}_{Control,\ Before} &= \hat{A} = \text{sample mean of } y \text{ for control group before policy implementation} \\ \bar{y}_{Treatment,\ Before} &= \hat{B} = \text{sample mean of } y \text{ for treatment group before policy implementation} \\ \bar{y}_{Control,\ After} &= \hat{E} = \text{sample mean of } y \text{ for control group after policy implementation} \\ \bar{y}_{Treatment,\ After} &= \hat{C} = \text{sample mean of } y \text{ for treatment group after policy implementation}\end{aligned}$$

The estimator $\hat{\delta}$ is called a **differences-in-differences** (abbreviated as D-in-D, DD, or DID) estimator of the treatment effect.

The estimator $\hat{\delta}$ can be conveniently calculated using a simple regression. Define y_{it} to be the observed outcome for individual i in period t. Let $AFTER_t$ be an indicator variable that equals one in the period after the policy change ($t = 2$) and zero in the period before the policy change ($t = 1$). Let $TREAT_i$ be a dummy variable that equals one if individual i is in the treatment group and zero if the individual is in the control group. Consider the regression model

$$y_{it} = \beta_1 + \beta_2 TREAT_i + \beta_3 AFTER_t + \delta(TREAT_i \times AFTER_t) + e_{it} \tag{7.19}$$

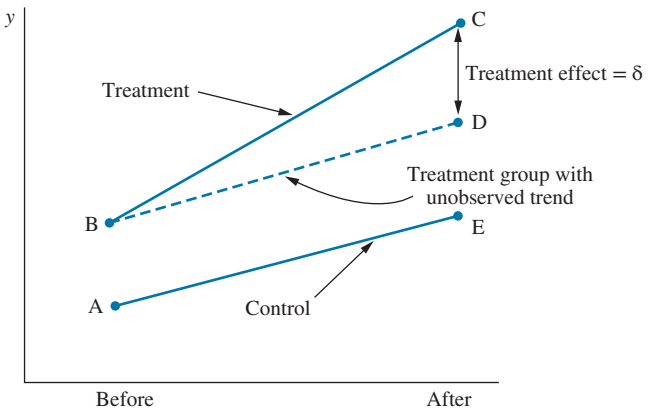


FIGURE 7.3 **Difference-in-Differences Estimation.**

The regression function is

$$E(y_{it}|\mathbf{X}) = \begin{cases} \beta_1 & TREAT = 0,\ AFTER = 0 \text{ [Control before = A]} \\ \beta_1 + \beta_2 & TREAT = 1,\ AFTER = 0 \text{ [Treatment before = B]} \\ \beta_1 + \beta_3 & TREAT = 0,\ AFTER = 1 \text{ [Control after = E]} \\ \beta_1 + \beta_2 + \beta_3 + \delta & TREAT = 1,\ AFTER = 1 \text{ [Treatment after = C]} \end{cases}$$

Here $\mathbf{X}$ contains the variables on the right-hand side of equation (7.19). In Figure 7.3, points $A = \beta_1, B = \beta_1 + \beta_2, E = \beta_1 + \beta_3$ and $C = \beta_1 + \beta_2 + \beta_3 + \delta$. Then

$$\begin{aligned} \delta &= (C - E) - (B - A) \\ &= \left[(\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3)\right] - \left[(\beta_1 + \beta_2) - \beta_1\right] \end{aligned}$$

Using this the least squares estimates b_1, b_2, b_3 and $\hat{\delta}$ from (7.19), we have

$$\begin{aligned} \hat{\delta} &= \left[\left(b_1 + b_2 + b_3 + \hat{\delta}\right) - (b_1 + b_3)\right] - \left[(b_1 + b_2) - b_1\right] \\ &= \left(\bar{y}_{Treatment,\ After} - \bar{y}_{Control,\ After}\right) - \left(\bar{y}_{Treatment,Before} - \bar{y}_{Control,Before}\right) \end{aligned}$$

EXAMPLE 7.12 | Estimating the Effect of a Minimum Wage Change: The DID Estimator

Card and Krueger (1994)[17] provide an example of a natural experiment and the **differences-in-differences estimator**. On April 1, 1992, New Jersey's minimum wage was increased from \$4.25 to \$5.05 per hour, while the minimum wage in Pennsylvania stayed at \$4.25 per hour. Card and Krueger collected data on 410 fast-food restaurants in New Jersey (the treatment group) and eastern Pennsylvania (the control group). These two groups are similar economically and close geographically, separated by only a river with multiple bridges. The "before" period is February 1992, and the "after" period is November 1992. Using these data, they estimate the effect of the "treatment," raising the New Jersey minimum wage on employment at fast-food restaurants in New Jersey. Their interesting finding, that there was no significant reduction[18] in employment, sparked a great debate and much further research.[19] In model (7.19), we will test the null and alternative hypotheses

$$H_0 : \delta \geq 0 \text{ versus } H_1 : \delta < 0 \qquad (7.20)$$

The relevant Card and Krueger data is in the data file *njmin3*. We use the sample means of *FTE*, the number of full-time equivalent[20] employees, given in Table 7.8, to estimate the treatment effect δ using the differences-in-differences estimator.

TABLE 7.8 Full-time Equivalent Employees by State and Period

Variable	*N*	Mean	se
Pennsylvania (PA)			
Before	77	23.3312	1.3511
After	77	21.1656	0.9432
New Jersey (NJ)			
Before	321	20.4394	0.5083
After	319	21.0274	0.5203

In Pennsylvania, the control group, employment fell during the period February to November. Recall that the

[17] David Card and Alan Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 316–361. We thank David Card for letting us use the data.

[18] Remember that failure to reject a null hypothesis does not make it true!

[19] The issue is hotly contested and the literature extensive. See, for example, http://en.wikipedia.org/wiki/Minimum_wage, and the references listed, as a starting point.

[20] Card and Krueger calculate FTE = 0.5 × number of part time workers + number of full time workers + number of managers.

minimum wage level was changed in New Jersey, but not in Pennsylvania, so that employment levels in Pennsylvania were not affected. In New Jersey we see an increase in *FTE* in the same period. The differences-in-differences estimate of the change in employment due to the change in the minimum wage is

$$\begin{aligned}\hat{\delta} &= \left(\overline{FTE}_{NJ,After} - \overline{FTE}_{PA,After}\right) - \left(\overline{FTE}_{NJ,Before} - \overline{FTE}_{PA,Before}\right) \\ &= (21.0274 - 21.1656) - (20.4394 - 23.3312) \\ &= 2.7536 \end{aligned} \tag{7.21}$$

We estimate that *FTE* employment increased by 2.75 employees during the period in which the New Jersey minimum wage was increased. This positive effect is contrary to what is predicted by economic theory.

Rather than compute the differences-in-differences estimate using sample means, it is easier and more general to use the regression format. In (7.19) let $y = FTE$ employment, the treatment variable is the indicator variable $NJ = 1$ if observation is from New Jersey, and zero if from Pennsylvania. The time indicator is $D = 1$ if the observation is from November and zero if it is from February. The differences-in-differences regression is then

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta(NJ_i \times D_t) + e_{it} \tag{7.22}$$

Using the 794 complete observations in the file *njmin3*, the least squares estimates are reported in column (1) of Table 7.9. At the $\alpha = 0.05$ level of significance the rejection region for the left-tail test in (7.20) is $t \le -1.645$, so we fail to reject the null hypothesis. We cannot conclude that the increase in the minimum wage in New Jersey reduced employment at New Jersey fast-food restaurants.

As with randomized control experiments, it is interesting to see the robustness of these results. In Table 7.9 column (2), we add indicator variables for fast-food chain and whether the restaurant was company-owned rather than franchise-owned. In column (3) we add indicator variables for geographical regions within the survey area. None of these changes alter the differences-in-differences estimate, and none lead to rejection of the null hypothesis in (7.20).

TABLE 7.9 Difference-in-Differences Regressions

	(1)	(2)	(3)
C	23.3312***	25.9512***	25.3205***
	(1.072)	(1.038)	(1.211)
NJ	−2.8918*	−2.3766*	−0.9080
	(1.194)	(1.079)	(1.272)
D	−2.1656	−2.2236	−2.2119
	(1.516)	(1.368)	(1.349)
D_NJ	2.7536	2.8451	2.8149
	(1.688)	(1.523)	(1.502)
KFC		−10.4534***	−10.0580***
		(0.849)	(0.845)
ROYS		−1.6250	−1.6934*
		(0.860)	(0.859)
WENDYS		−1.0637	−1.0650
		(0.929)	(0.921)
CO_OWNED		−1.1685	−0.7163
		(0.716)	(0.719)
SOUTHJ			−3.7018***
			(0.780)
CENTRALJ			0.0079
			(0.897)
PA1			0.9239
			(1.385)
N	794	794	794
R^2	0.007	0.196	0.221
adj. R^2	0.004	0.189	0.211

Standard errors in parentheses
Two-tail *p*-values: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

EXAMPLE 7.13 | Estimating the Effect of a Minimum Wage Change: Using Panel Data

In the previous section's differences-in-differences analysis, we did not exploit one very important feature of Card and Krueger's data—namely, that the *same* fast-food restaurants were observed on two occasions. We have "before" and "after" data on 384 of the 410 restaurants. These are called **paired data** observations, or **repeat data** observations, or **panel data** observations. In Chapter 1 we introduced the notion of a **panel** of data—we observe the same individual-level units over several periods. The Card and Krueger data includes $T = 2$ observations on $N = 384$ individual restaurants among the 410 restaurants surveyed. The remaining 26 restaurants had missing data on *FTE*

either in the "before" or "after" period. There are powerful advantages to using panel data, some of which we will describe here. See Chapter 15 for a much more extensive discussion.

Using panel data, we can control for **unobserved individual-specific characteristics**. There are characteristics of the restaurants that we do not observe. Some restaurants will have preferred locations, some may have superior managers, and so on. These unobserved individual specific characteristics are included in the error term of the regression (7.22). Let c_i denote any unobserved characteristics of individual restaurant i that do not change over time. Adding c_i to (7.22), we have

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta(NJ_i \times D_t) + c_i + e_{it} \quad (7.23)$$

Whatever c_i might be, it contaminates this regression model. A solution is at hand *if* we have a panel of data. If we have $T = 2$ repeat observations, we can *eliminate* c_i by analyzing the changes in *FTE* from period one to period two. Recall that $D_t = 0$ in period one, so $D_1 = 0$; and $D_t = 1$ in period two, so $D_2 = 1$. Subtract the observation for $t = 1$ from that for $t = 2$

$$\begin{aligned} FTE_{i2} &= \beta_1 + \beta_2 NJ_i + \beta_3 1 + \delta(NJ_i \times 1) + c_i + e_{i2} \\ -(FTE_{i1} &= \beta_1 + \beta_2 NJ_i + \beta_3 0 + \delta(NJ_i \times 0) + c_i + e_{i1}) \\ \hline \Delta FTE_i &= \beta_3 + \delta NJ_i + \Delta e_i \end{aligned}$$

where $\Delta FTE_i = FTE_{i2} - FTE_{i1}$ and $\Delta e_i = e_{i2} - e_{i1}$. Using the **differenced data**, the regression model of interest becomes

$$\Delta FTE_i = \beta_3 + \delta NJ_i + \Delta e_i \quad (7.24)$$

Observe that the contaminating factor c_i has dropped out! Whatever those unobservable features might have been, they are now gone. The intercept β_1 and the coefficient β_2 have also dropped out, with the parameter β_3 becoming the new intercept. The most important parameter, δ, measuring the treatment effect is the coefficient of the indicator variable NJ_i, which identifies the treatment (New Jersey) and control group (Pennsylvania) observations.

The estimated model (7.24) is

$$\begin{aligned} \widehat{\Delta FTE} &= -2.2833 + 2.7500 NJ \quad R^2 = 0.0146 \\ (\text{se}) \quad & (1.036) \quad (1.154) \end{aligned}$$

The estimate of the treatment effect $\hat{\delta} = 2.75$ using the differenced data, which accounts for any unobserved individual differences, is very close to the differences-in-differences estimate. Once again we fail to conclude that the minimum wage increase has reduced employment in these New Jersey fast-food restaurants.

7.6 Treatment Effects and Causal Modeling

In Section 7.5, we provided the basics of treatment effect models. In this section, we present extensions and enhancements using the framework of **potential outcomes**, sometimes called the **Rubin Causal Model (RCM)**, in recognition of Donald B. Rubin who formulated this approach.[21]

7.6.1 The Nature of Causal Effects

Economists are interested in causal relationships between variables. **Causality**, or causation, means that a change in one variable is the direct consequence of a change in another variable. For example, if you receive an hourly wage rate, then increasing your work hours (the cause) will lead to an increase in your income (the effect). Another example is from the standard supply and demand model for a normal good. If consumer incomes rise (the cause), demand increases, and there is a subsequent increase in the market price and quantities bought and sold (the effect).

A cause must precede, or be contemporaneous with, the effect. The confusion between correlation and causation is widespread, and correlation does not imply causation. We observe many associations between variables that are not causal. The correlation between the divorce rate in

[21] The literature in this area has grown dramatically in recent years, and continues to grow. In this section we draw heavily on a survey by Guido W. Imbens and Jeffrey M. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86, Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21; and Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. These references are advanced. See also Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.

the state of Maine and the U.S. per capita consumption of margarine is 0.9926[22] over the period 2000–2009. We doubt that this high correlation is a causal relationship. Not all confusions, or spurious correlations, are amusing and harmless. There is a concern among some parents about the relationship between childhood vaccinations and subsequent negative health outcomes, such as autism. Despite intense study by the U.S. Centers for Disease Control and Prevention (CDC), finding no causal relationship, there has been a movement among parents to not have some vaccinations for their children, resulting in concern by health officials that some childhood diseases will make a widespread comeback.

7.6.2 Treatment Effect Models

Treatment effect models seek to estimate a causal effect. Let the treatment, which might be an individual receiving a new drug, or some additional job training, be denoted as $d_i = 1$, whereas not receiving the treatment is $d_i = 0$. The outcome of interest might be a cholesterol level if the treatment is a new drug. If the treatment is job training, the outcome might be a worker's performance on completing a particular task. For each individual there are two possible, or potential, outcomes, y_{1i} if an individual receives treatment $(d_i = 1)$, and y_{0i} if the individual does not receive treatment $(d_i = 0)$. We would like to know the **causal effect** $y_{1i} - y_{0i}$, the difference in the outcome for individual i if they receive the treatment versus if they do not. An advantage of the potential outcomes framework is that it forces us to recognize that the treatment effect varies across individuals—it is individual specific. The difficulty is that we never observe both y_{1i} and y_{0i}. We only observe one or the other. The outcome we observe is

$$y_i = \begin{cases} y_{1i} & \text{if } d_i = 1 \\ y_{0i} & \text{if } d_i = 0 \end{cases} \tag{7.25}$$

Written another way, what we observe is

$$y_i = y_{1i}d_i + y_{0i}(1 - d_i) = y_{0i} + (y_{1i} - y_{0i})\, d_i \tag{7.26}$$

Instead of being able to estimate $y_{1i} - y_{0i}$ for each individual, what we are able to estimate is the population **average treatment effect (ATE)**, $\tau_{ATE} = E(y_{1i} - y_{0i})$. To see this, express the difference between the conditional expectation of y_i, the outcome we actually observe, for those who receive treatment, $(d_i = 1)$, and those who do not, $(d_i = 0)$;

$$E(y_i|d_i = 1) - E(y_i|d_i = 0) = E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 0) \tag{7.27}$$

In a randomized, controlled experiment, individuals are randomly selected from the population and then randomly assigned to a group receiving the treatment (the **treatment group**), for whom $(d_i = 1)$, or to a group not receiving the treatment (the **control group**), for whom $(d_i = 0)$. In this way the treatment, d_i, is statistically independent of the potential outcomes y_{1i} and y_{0i} so that

$$\begin{aligned} E(y_i|d_i = 1) - E(y_i|d_i = 0) &= E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 0) \\ &= E(y_{1i}) - E(y_{0i}) = E(y_{1i} - y_{0i}) \\ &= \tau_{ATE} \end{aligned} \tag{7.28}$$

From the first line to the second we use the fact that if two random variables, say X and Y, are statistically independent,[23] then $E(Y|X = x) = E(Y)$. To see that this is true, suppose X and Y are discrete random variables. Then

$$E(Y) = \sum yP(Y = y) \text{ and } E(Y|X = x) = \sum yP(Y = y|X = x)$$

[22] http://www.tylervigen.com/

[23] We revert to the notation from the probability primer here, with upper case Y and X being random variables and lower case y and x being values of the random variables.

If X and Y are statistically independent, then

$$P(Y = y|X = x) = P(Y = y)$$

so that

$$E(Y|X = x) = \sum yP(Y = y|X = x) = \sum yP(Y = y) = E(Y)$$

If we randomly choose population members and randomly assign them to the treatment and control groups, then treatment, d_i, is statistically independent of the potential outcomes of the experiment. An unbiased estimator of $E(y_i|d_i = 1)$ is the sample mean of the N_1 outcomes for the treatment group, $\bar{y}_1 = \sum_{i=1}^{N_1} y_{1i}/N_1$. An unbiased estimator of $E(y_i|d_i = 0)$ is the sample mean of the N_0 outcomes for the control group, $\bar{y}_0 = \sum_{i=1}^{N_0} y_{0i}/N_0$. An unbiased estimator of the population average treatment effect is $\hat{\tau}_{ATE} = \bar{y}_1 - \bar{y}_0$. This is the **difference estimator** in equation (7.14). That is, we can obtain the estimator of the average treatment effect from the simple regression $y_i = \alpha + \tau_{ATE} d_i + e_i$ using all $N = N_0 + N_1$ observations.

7.6.3 Decomposing the Treatment Effect

Using equation (7.27) $\left[E(y_i|d_i = 1) - E(y_i|d_i = 0) = E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 0)\right]$, we can gain additional insight into the simple regression $y_i = \alpha + \tau_{ATE} d_i + e_i$. Add and subtract $E(y_{0i}|d_i = 1)$ to the right-hand side, and rearrange to obtain

$$\begin{aligned} E(y_i|d_i = 1) - E(y_i|d_i = 0) = &\left[E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)\right] \\ &+ \left[E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)\right] \end{aligned} \tag{7.29}$$

The left-hand side is the difference in average outcomes for the treatment group $(d_i = 1)$ and the control group $(d_i = 0)$. The difference $\left[E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)\right]$ is average difference in potential outcomes for those who received the treatment, or as called in this literature, the **average treatment effect on the treated (ATT)**, which we denote by τ_{ATT}. The second term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ is the average potential outcome for those in the treatment group should they not receive treatment minus the average outcome for those in the control group. If individuals are truly randomly assigned to treatment and control groups $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ will be zero, meaning that there are no differences between the expected potential outcomes for the treatment and control groups if they had remained untreated. In this case, the treatment effect $\tau_{ATE} = E(y_i|d_i = 1) - E(y_i|d_i = 0)$ equals $\tau_{ATT} = E(y_{1i}|d_i = 1) - E(y_{0i}|d_i = 1)$, the average treatment effect on the treated.

In equation (7.29), if the second term in brackets is not zero, or $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) \neq 0$, then there is **selection bias**. It means that individuals are not randomly assigned to the treatment and control groups because the average of the potential outcomes if untreated, y_{0i}, in the treatment and control groups are different. If the treatment is receiving a new drug, there is selection bias if (i) a screener looks at a randomly chosen person and thinks "This person looks sickly and could use this drug, so I'll assign him to the treatment group;" or (ii) a person thinks the treatment might be good for him, and manages to be added to the treatment group. Either way, there is a difference in the average untreated health y_{0i} of the treatment and control groups. The term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0)$ is called **selection bias** for this reason. Random assignment of individuals to treatment and control groups eliminates selection bias. If there is selection bias, then the difference estimator $\hat{\tau}_{ATE} = \bar{y}_1 - \bar{y}_0$ is not an unbiased estimator of the average treatment effect, and the average treatment effect is not the average treatment effect on the treated.

To summarize, in a randomized experiment the treatment indicator d_i is statistically independent of the potential outcomes y_{0i} and y_{1i}. We do not observe both potential outcomes but rather $y_i = y_{0i} + (y_{1i} - y_{0i})d_i$. If treatment d_i is statistically independent of the potential outcomes, then

$$\tau_{ATE} = \tau_{ATT} = E(y_i|d_i = 1) - E(y_i|d_i = 0) \tag{7.30}$$

and an unbiased estimator is

$$\hat{\tau}_{ATE} = \hat{\tau}_{ATT} = \bar{y}_1 - \bar{y}_0 \tag{7.31}$$

The equality $\tau_{ATE} = \tau_{ATT}$ actually holds under a weaker assumption than statistical independence. From (7.29)

$$\tau_{ATE} = \tau_{ATT} + E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) \tag{7.32}$$

The selection bias term $E(y_{0i}|d_i = 1) - E(y_{0i}|d_i = 0) = 0$ if $E(y_{0i}|d_i = 1) = E(y_{i0})$ and $E(y_{0i}|d_i = 0) = E(y_{i0})$. This is called the **conditional independence assumption (CIA)**, or **conditional mean independence**. While this is a less stringent condition than statistical independence between the treatment and the potential outcomes, it is still strong. It suggests that being in the treatment or control group is unrelated to the average outcome for the untreated.

7.6.4 Introducing Control Variables

A **control variable**, x_i, is not the object of interest in a study. It is included in the model to hold constant factors that, if neglected, would lead to selection bias. See Section 6.3.4. In treatment effect models, control variables are introduced in order to allow unbiased estimation of the treatment effect when the potential outcomes, y_{0i} and y_{1i}, might be correlated with the treatment variable, d_i. Ideally, by conditioning on a control variable x_i the treatment becomes "as good as" randomized, allowing us to estimate the average causal or treatment effect. We consider only a single control variable to simplify our presentation. The methods discussed as follows carry over to the case with multiple control variables. The key is an extension of the **conditional independence assumption**,[24]

$$E(y_{0i}|d_i, x_i) = E(y_{0i}|x_i) \quad \text{and} \quad E(y_{1i}|d_i, x_i) = E(y_{1i}|x_i) \tag{7.33}$$

Once we condition on the control variables, then the expected potential outcomes do not depend the treatment. In a sense, having good control variables *is as good as* having a randomized controlled experiment. Good control variables have the feature of being "predetermined" in the sense that they are fixed, and given, at the time the treatment is assigned. Enough control variables should be added so that the conditional independence assumption holds. Avoid "bad control" variables that might be outcomes of the treatment.

When potential outcomes depend on x_i, then the average treatment effect depends on x_i, and is

$$\tau_{ATE}(x_i) = E(y_{1i}|d_i, x_i) - E(y_{0i}|d_i, x_i) = E(y_{1i}|x_i) - E(y_{0i}|x_i)$$

Assuming a linear regression structure for the expectations, and recalling that the observed outcome is $y_i = y_{0i} + (y_{1i} - y_{0i})\, d_i$, let

$$E(y_i|x_i, d_i = 0) = E(y_{i0}|x_i, d_i = 0) = E(y_{i0}|x_i) = \alpha_0 + \beta_0 x_i \tag{7.34a}$$

$$E(y_i|x_i, d_i = 1) = E(y_{i1}|x_i, d_i = 1) = E(y_{i1}|x_i) = \alpha_1 + \beta_1 x_i \tag{7.34b}$$

The treatment effect is the difference between equations (7.34b) and (7.34a), or

$$\tau_{ATE}(x_i) = (\alpha_1 + \beta_1 x_i) - (\alpha_0 + \beta_0 x_i) = (\alpha_1 - \alpha_0) - (\beta_1 - \beta_0) x_i \tag{7.35}$$

Because $\tau_{ATE}(x_i)$ depends on x_i, the average treatment effect will be obtained by "averaging" over the population distribution of x_i. Recall from the probability primer that a "population average"

[24] This assumption has been called *unconfoundedness* and also *ignorability*. The literature on causal modeling spans several disciplines, and the terminology can be quite different in each. The following development follows Woodridge (2010, 919–920).

is an expected value. So we define the average treatment effect as $\tau_{ATE} = E_x\left[\tau_{ATE}(x_i)\right]$ where the subscript x on the expectation operator means that we are treating x as random.

In practice, we can estimate the regression functions separately on the treatment and control groups:

1. Obtain $\hat{\alpha}_0 + \hat{\beta}_0 x_i$ from a regression of y_i on x_i for the control group, $(d_i = 0)$
2. Obtain $\hat{\alpha}_1 + \hat{\beta}_1 x_i$ from a regression of y_i on x_i for the treatment group, $(d_i = 1)$

Then

$$\hat{\tau}_{ATE}(x_i) = \hat{\alpha}_1 + \hat{\beta}_1 x_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 x_i\right) = \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)x_i \tag{7.36}$$

Averaging the estimated value across the sample values gives

$$\begin{aligned}\hat{\tau}_{ATE} &= N^{-1}\sum_{i=1}^{N}\hat{\tau}_{ATE}(x_i) = N^{-1}\sum_{i=1}^{N}\left[\left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)x_i\right] \\ &= \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\left(N^{-1}\sum_{i=1}^{N} x_i\right) \\ &= \left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\beta}_1 - \hat{\beta}_0\right)\bar{x}\end{aligned} \tag{7.37}$$

Using slope and intercept indicator variables, we can estimate the average treatment effect in a pooled regression, and calculate a standard error for the estimate $\hat{\tau}_{ATE}$. The pooled regression is

$$y_i = \alpha + \theta d_i + \beta x_i + \gamma(d_i x_i) + e_i \tag{7.38}$$

The regression functions for the treatment and control groups are

$$E(y_i|d_i, x_i) = \begin{cases} \alpha + \beta x_i & \text{if } d_i = 0 \\ (\alpha + \theta) + (\beta + \gamma)x_i & \text{if } d_i = 1 \end{cases} \tag{7.39}$$

In terms of the separate regression coefficients

$$\alpha = \alpha_0, \quad \beta = \beta_0, \quad \alpha + \theta = \alpha_1, \quad \text{and} \quad \beta + \gamma = \beta_1 \tag{7.40}$$

It follows that from the pooled regression (7.38) the estimates $\hat{\theta} = \hat{\alpha}_1 - \hat{\alpha}_0$ and $\hat{\gamma} = \hat{\beta}_1 - \hat{\beta}_0$. The relation of these estimates to $\hat{\tau}_{ATE}$ is

$$\hat{\theta} = \hat{\tau}_{ATE} - \bar{x}\left(\hat{\beta}_1 - \hat{\beta}_0\right) = \hat{\tau}_{ATE} - \bar{x}\hat{\gamma}$$

or

$$\hat{\tau}_{ATE} = \hat{\theta} + \bar{x}\hat{\gamma}$$

We can modify the pooled regression so that τ_{ATE} appears in the pooled regression. In the pooled regression (7.38) add and subtract the term $\gamma(d_i\bar{x})$

$$\begin{aligned} y_i &= \alpha + \theta d_i + \beta x_i + \gamma(d_i x_i) + \left[\gamma d_i\bar{x} - \gamma d_i\bar{x}\right] + e_i \\ &= \alpha + \left(\theta + \gamma\bar{x}\right)d_i + \beta x_i + \gamma\left[d_i(x_i - \bar{x})\right] + e_i \\ &= \alpha + \tau_{ATE} d_i + \beta x_i + \gamma(d_i\tilde{x}_i) + e_i \end{aligned} \tag{7.41}$$

Now the population average treatment effect τ_{ATE} is a parameter in the pooled regression. The term $\tilde{x}_i = (x_i - \bar{x})$ is notation for deviations about the mean. By using least squares regression, we obtain $\hat{\tau}_{ATE}$. Your software will also report a standard error $\text{se}(\hat{\tau}_{ATE})$.[25]

The average treatment effect in the population, $\tau_{ATE} = E(y_{1i} - y_{0i})$, may not be the parameter of interest in some applications. By slightly modifying the pooled regression, we can obtain the

[25] Wooldridge (2010, p. 919) notes that the usual estimator of the standard error is not quite valid in this case because it ignores the additional variability added by including the sample mean in $\tilde{x}_i = (x_i - \bar{x})$. One alternative to the usual standard error is to use the **bootstrap** standard error, discussed in Appendix 5B.5.

average treatment effect of a subpopulation. For example, how large is the average treatment effect on those who actually received treatment? The **average treatment effect on the treated**, τ_{ATT}, where the subscript *ATT* denotes the target group, is obtained by estimating the pooled regression

$$y_i = \alpha + \tau_{ATT} d_i + \beta x_i + \gamma\left(d_i \tilde{x}_{i1}\right) + e_i \tag{7.42}$$

where $\tilde{x}_{i1} = \left(x_i - \bar{x}_1\right)$ and $\bar{x}_1 = N_1^{-1} \sum_{i=1}^{N_1} x_i$ for the treatment group, where $d_i = 1$.

Similarly, we can restrict measurement of the treatment effect to other subpopulations of interest. For example, if we are considering the effects of a job training program, we may not want to include the extremely wealthy. We could specify the population of interest to be those with incomes in the lowest 25% of society. Denote this restricted group of interest by R and let $\tau_{ATE,\,R}$ be the average treatment effect on this group. Let $\tilde{x}_{iR} = \left(x_i - \bar{x}_R\right)$, where $\bar{x}_R = N_R^{-1} \sum_{i \in R} x_i$, with $i \in R$ indicating that we are restricting the sum to those individuals i falling in the target group, R, and N_R is the number of individuals in the sample satisfying the condition. Then we can estimate $\tau_{ATE,R}$ from the pooled regression

$$y_i = \alpha + \tau_{ATE,R} d_i + \beta x_i + \gamma\left(d_i \tilde{x}_{iR}\right) + e_i \tag{7.43}$$

7.6.5 The Overlap Assumption

The so-called **overlap assumption** must hold, in addition to the conditional independence assumption in equation (7.33). The overlap assumption says that for each value of x_i it must be possible to see an individual in the treatment and control groups, or $0 < P\left(d_i = 1 | x_i\right) < 1$ and $0 < P\left(d_i = 0 | x_i\right) = 1 - P\left(d_i = 1 | x_i\right) < 1$. A rule of thumb is to compute the normalized difference

$$\frac{\bar{x}_1 - \bar{x}_0}{\left(s_1^2 + s_0^2\right)^{1/2}} \tag{7.44}$$

where s_1^2 and s_0^2 are the sample variances of the explanatory variable x for the treatment and control groups. If the normalized difference is greater in absolute value than 0.25,[26] then there is cause for concern. If the overlap assumption fails, then redefining the population of interest may be required. To see the impact of the difference of means, $\bar{x}_1 - \bar{x}_0$, on the average treatment effect, let $f_0 = N_0/N$ and $f_1 = N_1/N$ be the fractions of observations in the control and treatment groups, respectively. In Appendix 7C, we show that

$$\hat{\tau}_{ATE} = \left(\bar{y}_1 - \bar{y}_0\right) - \left(f_0 \hat{\beta}_1 + f_1 \hat{\beta}_0\right)\left(\bar{x}_1 - \bar{x}_0\right)$$

If the difference in the sample means of the treatment and control groups is large, the estimated slopes from the regressions in (7.34), $\hat{\beta}_1$ and $\hat{\beta}_0$, have a larger influence in the estimate $\hat{\tau}_{ATE}$ of the average treatment effect.

7.6.6 Regression Discontinuity Designs

Regression discontinuity (RD) designs[27] arise when the separation into treatment and control groups follows a deterministic rule, such as "Students receiving 75% or higher on the midterm exam will receive an award." How the award affects future academic outcomes might be the question of interest. The key insight about the RD designs is that that students receiving "close to

[26] Wooldridge (2010, p. 917)

[27] In this section we draw heavily on a survey by David S. Lee and Thomas Lemieux (2010) "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(1), 5-86, Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21 and Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Chapter 6. These references are advanced. See also Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press, Chapter 4.

75%" are likely very similar in most regards (a condition that can be checked) so that those just below the cutoff point are a good comparison group for those just above the cutoff. Using individuals close to the cutoff is "just as good as" a random assignment, for the purpose of estimating a treatment effect.

Suppose that x_i is the single variable determining whether an individual is assigned to the treatment group or control group. In this literature, x_i is called the **forcing variable**. The treatment indicator variable $d_i = 1$ if $x_i \geq c$, where c is a preassigned cutoff value and $d_i = 0$ if $x_i < c$. This is said to be a **sharp regression discontinuity design** because the treatment is definitely given if the forcing variable crosses the threshold. The observed outcome is $y_i = (1 - d_i)y_{0i} + d_i y_{1i}$, where y_{0i} is the potential outcome for individual i when not receiving treatment and y_{1i} is the potential outcome for individual i when receiving the treatment. For the sharp RD design, the conditional independence assumption in equation (7.33)

$$E(y_{0i}|d_i, x_i) = E(y_{0i}|x_i) \quad \text{and} \quad E(y_{1i}|d_i, x_i) = E(y_{1i}|x_i)$$

is automatically satisfied because the treatment is completely determined by the forcing variable, x_i. Interestingly, the overlap assumption fails completely. For a given value of x_i, we cannot hope to observe individuals in both treatment and control groups. Rather than trying to estimate a population average treatment effect, in the RD design we estimate the treatment effect "at the cutoff,"

$$\tau_c = E(y_{1i} - y_{0i}|x_i = c) = E(y_{1i}|x_i = c) - E(y_{0i}|x_i = c) \tag{7.45}$$

One required assumption is "continuity." That is, $E(y_{1i}|x_i)$ and $E(y_{0i}|x_i)$ must meet smoothly at $x_i = c$ except for a "jump." The jump is the treatment effect at the cutoff, τ_c.

A picture is worth a thousand words, especially with RD designs, so let us look at a graph. Suppose we give a 100 point midterm exam (the forcing variable x) and award a new laptop computer to students receiving a score of 75 (the cutoff value c) or over. The outcome we measure is student performance, y, on a 400 point final exam.

In Figure 7.4, based on simulated data, we see that at midterm score 75 there is a jump in the final exam score. That jump is what we seek to measure. The RDD idea is that students receiving just under and just over 75 are basically very similar, so that if we compare them it is just as good as randomly assigning treatment. Another way to picture the outcomes is to divide the forcing

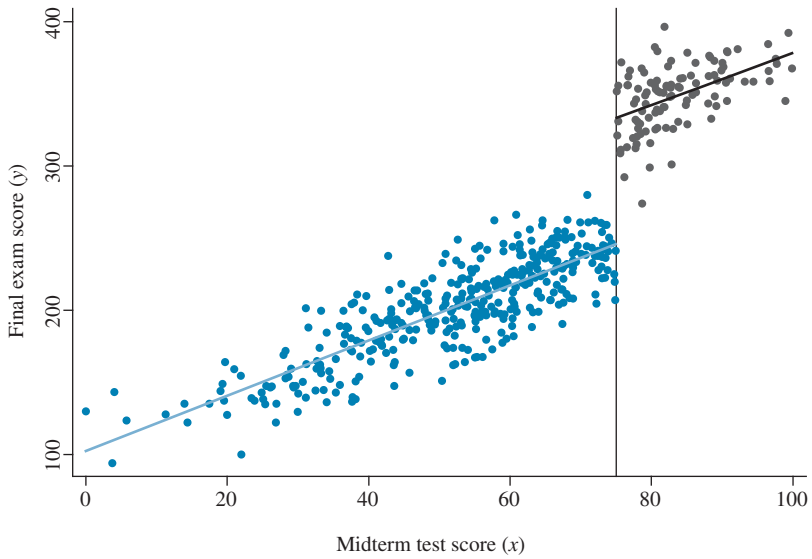


FIGURE 7.4 **Regression Discontinuity Design.**

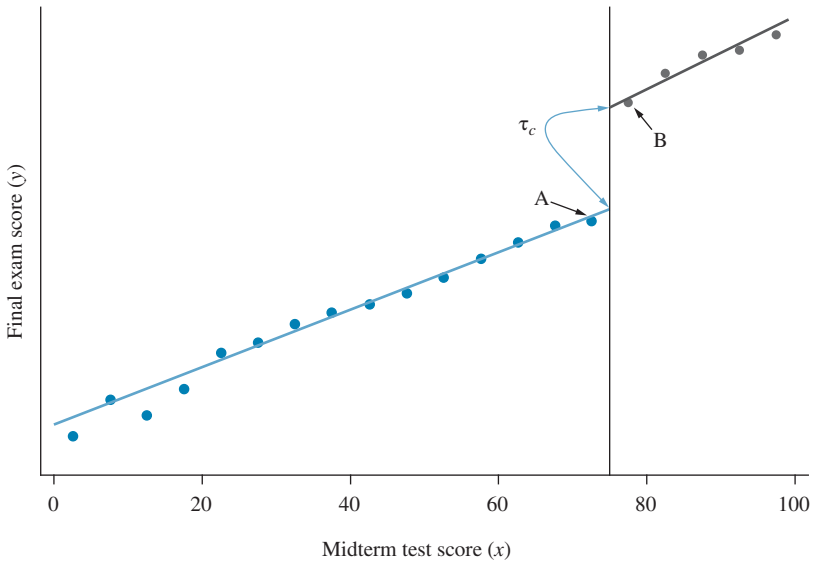


FIGURE 7.5 **Conditional Means Graph.**

variable (x) into intervals, or bins, and calculate and plot the mean, or median, of the outcome variable (y). Figure 7.5 is based on five point bins.

The difference between the mean scores of the two groups (A and B) just to either side of the cutoff is an estimate of the treatment effect at the cutoff, in this case $\hat{\tau}_c = \text{B} - \text{A} = 326.7 - 243.6 = 83.1$. We estimate that for students near the cutoff, getting a 75 or higher on the midterm, and thus receiving a new computer, had scores on the final exam that were 83.1 points higher than those who were also near the cutoff, but not receiving the prize, all other things being equal. This estimator is reasonable and intuitive. The difficulty is that students in the 70–75 range of test scores may not be as similar as we would like to students with test scores 75–80. If we make the bin widths smaller and smaller, then the groups to either side of the cutoff become more and more similar, but the number of observations in each bin gets smaller and smaller, reducing the reliability of this estimator of the treatment effect.[28]

Instead, let us use all the observations and use regression analysis to estimate the treatment effect at the cutoff, τ_c. Estimate the regression functions separately on the two groups, using as explanatory variable $x_i - c$:

1. Obtain $\hat{\alpha}_0 + \hat{\beta}_0(x_i - c)$ from a regression of y_i on $x_i - c$ for individuals below the cutoff, $(x_i < c)$.
2. Obtain $\hat{\alpha}_1 + \hat{\beta}_1(x_i - c)$ from a regression of y_i on $x_i - c$ for individuals above the cutoff, $(x_i \geq c)$.

The estimate of τ_c is $\hat{\tau}_c = \hat{\alpha}_1 - \hat{\alpha}_0$. Equivalently, we can use a pooled regression with an indicator variable. Define $d_i = 1$ if $x_i \geq c$, and $d_i = 0$ if $x_i < c$. Then the equivalent pooled regression is

$$y_i = \alpha + \tau_c d_i + \beta(x_i - c) + \gamma[d_i(x_i - c)] + e_i \tag{7.46}$$

There are some additional considerations when using RD designs. First, using the full range of the data may not be a good idea. The goal is to estimate the regression "jump" at the cutoff value

[28] Selecting bin width is an important issue in RDD analysis. See Lee and Lemieux (2010, pp. 307–314).

$x_i = c$. With sufficient observations, we can make the estimate "local" by only using data within a certain distance h of the cutoff. That is, use observations for which $c - h \le x_i \le c + h$. Checking the robustness of findings to various choices of h is a good idea.

Second, it is important to build into the regression sufficient flexibility to capture a nonlinear relationship. For example, if the true relationship between the outcome y and the test score x is nonlinear, then using linear relationships in the RDD can give a biased estimator of the treatment effect. In Figure 7.6, we illustrate a situation when there is no "jump" in the underlying relationship but using RDD with an assumed linear fit makes there appear be a positive treatment effect at $x_i = c$.

For this reason, researchers often use additional powers of $(x_i - c)$ in the regression relation, such as $(x_i - c)^2$, $(x_i - c)^3$, and $(x_i - c)^4$. If we use up to the third power, the pooled regression becomes

$$y_i = \alpha + \tau_c d_i + \sum_{q=1}^{3} \beta_q (x_i - c)^q + \sum_{p=1}^{3} \gamma_p \left[d_i (x_i - c)^p \right] + e_i \tag{7.47}$$

For the data in Figure 7.6, the estimated treatment effect from (7.47), $\hat{\tau}_c$, is not statistically different from zero, with a $t = 1.11$ and a p-value of 0.268. Alternatively, the recognition of a "nonjump" could be detected by using local observations for which $c - h \le x_i \le c + h$.

Third, it is possible that variables other than the forcing variable, say z_i, may influence the outcome. These can be added to the RDD model in equation (7.47).

Fourth, the illustration we have provided assumes that those with test scores at 75 or above are given a new computer whether they want one or not. We could instead offer those with test scores 75 and above a heavily discounted price on a new computer before the final exam. Some will elect to purchase the new machine using the discount and others will not. Some with test scores below 75 could, of course, also buy new computers. These issues lead to what is known as a **fuzzy regression discontinuity design**. The key in this case is that there is a "jump" in the **probability of treatment** (receiving a new computer before the final exam) at $x_i = c$. In this case, we must use an estimation alternative to least squares called **instrumental variables estimation**. This topic is considered in Chapter 10.

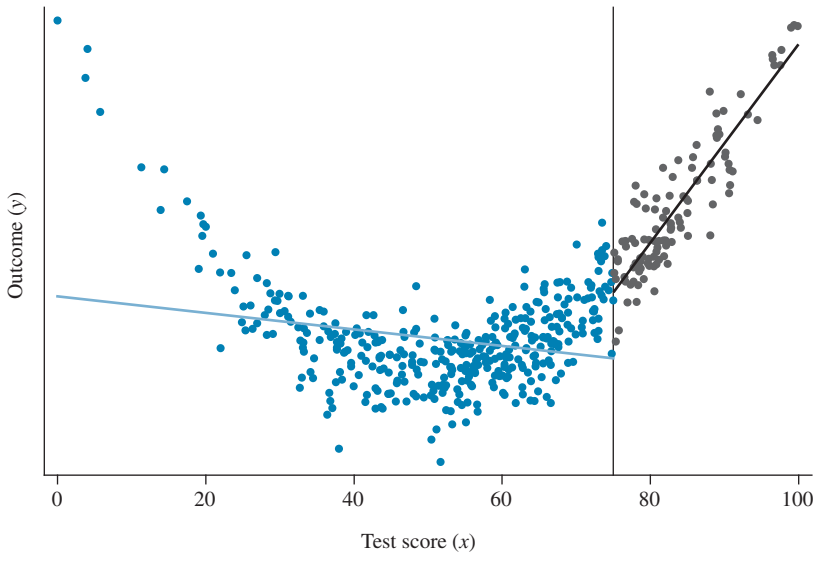


FIGURE 7.6 **RDD bias.**

7.7 Exercises

7.7.1 Problems

7.1 Suppose we are able to collect a random sample of data on economics majors at a large university. Further suppose that, for those entering the workforce, we observe their employment status and salary 5 years after graduation. Let SAL = \$ salary for those employed, GPA = grade point average on a 4.0 scale during their undergraduate program, with $METRICS = 1$ if student took econometrics, $METRICS = 0$ otherwise.

a. Consider the regression model $SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + e$. Should we consider this a causal model, or a predictive model? Explain your reasoning.

b. Assuming β_2 and β_3 are positive, draw a sketch of $E(SAL|GPA, METRICS) = \beta_1 + \beta_2 GPA + \beta_3 METRICS$.

c. Define a dummy variable $FEMALE = 1$, if the student is female; 0 otherwise. Modify the regression model to be $SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \delta_1 FEMALE + e$. What is the expected salary of a male who has not taken econometrics? What is the expected salary of a female who has taken econometrics?

d. Consider the regression model

$$SAL = \beta_1 + \beta_2 GPA + \beta_3 METRICS + \delta_1 FEMALE + \delta_2 (FEMALE \times METRICS) + e \quad \text{(XR7.1.1)}$$

What is the expected salary of a male who has not taken econometrics? What is the expected salary of a female who has taken econometrics?

e. In the equation (XR7.1.1), assume that $\delta_1 < 0$ and $\delta_2 < 0$. Sketch $E(SAL|GPA, METRICS, FEMALE)$ versus GPA for (i) males not taking econometrics, (ii) males taking econometrics, (iii) females not taking econometrics, and (iv) females taking econometrics.

f. In equation (XR7.1.1), what are the null and alternative hypotheses, in terms of model parameters, for testing that econometrics training does not affect the average salary of economics majors? In order to use the test statistic in equation (6.4), what regression must you estimate in addition to (XR7.1.1)? What is the distribution of the test statistic if the null hypothesis is true assuming $N = 300$? What is the rejection region for a 5% test?

7.2 In September of 1998, a local TV station contacted an econometrician to analyze some data for them. They were going to do a Halloween story on the legend of full moons affecting behavior in strange ways. They collected data from a local hospital on emergency room cases for the period from January 1, 1998 until mid-August. There were 229 observations. During this time, there were eight full moons and seven new moons (a related myth concerns new moons) and three holidays (New Year's day, Memorial Day, and Easter). If there is a full-moon effect, then hospital administrators will adjust numbers of emergency room doctors and nurses, and local police may change the number of officers on duty. Let T be a time trend ($T = 1, 2, 3, \ldots, 229$). Let the indicator variables $HOLIDAY = 1$ if the day is a holiday, $= 0$ otherwise; $FRIDAY = 1$ if the day is a Friday, $= 0$ otherwise; $SATURDAY = 1$ if the day is a Saturday, $= 0$ otherwise; $FULLMOON = 1$ if there is a full moon, $= 0$ otherwise; $NEWMOON = 1$ if there is a new moon, $= 0$ otherwise. Consider the model

$$CASES = \beta_1 + \beta_2 T + \delta_1 HOLIDAY + \delta_2 FRIDAY + \delta_3 SATURDAY + \theta_1 FULLMOON + \theta_2 NEWMOON + e \quad \text{(XR7.2.1)}$$

a. What is the expected number of emergency room cases for day $T = 100$, which was a Friday with neither a full or new moon?

b. What is the expected number of emergency room cases for day $T = 185$, which was a holiday Saturday?

c. In terms of the model parameters, what are the null and alternative hypotheses for testing that neither a full moon nor a new moon have any effect on the number of emergency room cases? What is the test statistic? What is the distribution of the test statistic if the null hypothesis is true? What is the rejection region for a 5% test?

d. The sum of squared residuals from the regression in (XR7.2.1) is 27109. If full moon and new moon are omitted from the model the sum of squared residuals is 27424. Carry out the test in (c). What is your conclusion?

e. Using the model in equation (XR7.2.1), the estimated coefficient of *SATURDAY* is 10.59 with standard error 2.12, and the estimated coefficient for *FRIDAY* is 6.91, with standard error 2.11. The estimated covariance between the coefficient estimators is 0.75. Should the hospitals prepare for significantly more emergency room patients on Saturday than Friday? State the relevant null and alternative hypotheses in terms of the model parameters. What is the test statistic? What is the distribution of the test statistic if the null hypothesis is true? What is the rejection region for a test at the 10% level? Carry out the test and state your conclusion?

7.3 One of the key problems regarding housing prices in a region concerns construction of "price indexes." That is, holding other factors constant, have prices increased, decreased or stayed relatively constant in a particular area? As an illustration, consider a regression model for house prices (in $1000s) on home sales from 1991 to 1996 in Stockton, CA, including as explanatory variables the size of the house (*SQFT*, in 100s of square feet), the age of the house (*AGE*) and annual indicator variables, such as $D92 = 1$ if the year is 1992 and 0 otherwise.

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \delta_1 D92 + \delta_2 D93 + \delta_3 D94 + \delta_4 D95 + \delta_5 D96 + e \quad \text{(XR7.3.1)}$$

An alternative model employs a "trend" variable $YEAR = 0, 1, \ldots, 5$ for the years 1991–1996.

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \tau YEAR + e \quad \text{(XR7.3.2)}$$

a. What is the expected selling price of a 10-year-old house with 2000 square feet of living space in each of the years 1991–1996 using equation (XR7.3.1)?

b. What is the expected selling price of a 10-year-old house with 2000 square feet of living space in each of the years 1991–1996 using equation (XR7.3.2)?

c. In order to choose between the models in (XR7.3.1) and (XR7.3.2), we propose a hypothesis test. What set of parameter constraints, or restrictions, would result in equation (XR7.3.1) equaling (XR7.3.2)? The sum of squared residuals from (XR7.3.1) is 2385745 and from (XR7.3.2) is 2387476. What is the test statistic for testing the restrictions that would make the two models equivalent? What is the distribution of the test statistic if the null hypotheses are true? What is the rejection region for a test at the 5% level? If the sample size is $N = 4682$, what do you conclude?

d. Using the model in (XR7.3.1) the estimated coefficients of the indicator variables for 1992 and 1994, and their standard errors, are −4.393 (1.271) and −13.174 (1.211), respectively. The estimated covariance between these two coefficient estimators is 0.87825. Test the null hypothesis that $\delta_3 = 3\delta_1$ against the alternative that $\delta_3 \neq 3\delta_1$ if $N = 4682$, at the 5% level.

e. The estimated value of τ in equation (XR7.3.2) is –4.12. What is the estimated difference in the expected house price for a 10-year-old house with 2000 square feet of living space in 1992 and 1994. Using information in (d), how does this compare to the result using (XR7.3.1)?

7.4 Angrist and Pischke[29] report estimation results of log-earnings equations using a large sample of college graduates. The predictors of interest (there are others included in their model) are the indicator variable *PRIVATE* (=1 if the individual attended a private college or university, = 0 if the individual attended a public college or university) and *SAT*/100, the individual's SAT score divided by 100. In the estimated regression equations, the dependent variable is ln(*EARNINGS*) and they include an intercept. The coefficient estimates, with standard errors in brackets, for two regressions that they estimate, are as follows.

$$0.212[0.060]PRIVATE \quad \text{(XR7.4.1)}$$

$$0.152\,[0.057]PRIVATE + 0.051[0.008](SAT/100) \quad \text{(XR7.4.2)}$$

a. In each model, what is the approximate effect on earnings of attending a private university rather than a public university?

[29] Joshua D. Angrist and Jörn-Steffen Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press, p. 66.

b. In the second model, what is the predicted effect on earnings of a 100-point increase in SAT score?

c. The estimated coefficient of *PRIVATE* is smaller in the second model than in the first model. Use the concept of "omitted variables bias" to explain this result.

d. What should happen to the estimated coefficients in equation (XR7.4.2) if parental income is included as an explanatory variable? Explain.

7.5 In 1985, the state of Tennessee carried out a statewide experiment with primary school students. Teachers and students were randomly assigned to be in a regular-sized class or a small class. The outcome of interest is a student's score on a math achievement test (*MATHSCORE*). Let $SMALL = 1$ if the student is in a small class and $SMALL = 0$ otherwise. The other variable of interest is the number of years of teacher experience, *TCHEXPER*.

a. Write down the econometric specification of the linear regression model explaining *MATHSCORE* as a function of *SMALL* and *TCHEXPER*. Use β_1, β_2, and β_3 as the model parameters. In this model, what is the expected math score for a child in a regular-sized class with a teacher having 10 years of experience? What is the expected math score for a child in a small class with a teacher having 10 years of experience?

b. Let $BOY = 1$ if the child is male and $BOY = 0$ if the child is female. Modify the model in part (a) to include the variables *BOY* and $BOY \times SMALL$, with parameters θ_1 and θ_2. Using this model

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the null hypothesis, written in terms of the model parameters, that the sex of the child has no effect on expected math score? What is the alternative hypothesis? What is the test statistic for the null hypothesis and what is its distribution if the null hypothesis is true? What is the test rejection region for a 5% test when $N = 1200$?

iv. It is conjectured that boys may benefit from small classes more than girls. What null and alternative hypothesis would you test to examine this conjecture? [*Hint*: Let the conjecture be the alternative hypothesis.]

7.6 In 1985, the state of Tennessee carried out a statewide experiment with primary school students. Teachers and students were randomly assigned to be in a regular-sized class or a small class. The outcome of interest is a student's score on a math achievement test (*MATHSCORE*). Let $SMALL = 1$ if the student is in a small class and $SMALL = 0$ otherwise. The other variable of interest is the number of years of teacher experience, *TCHEXPER*. Let $BOY = 1$ if the child is male and $BOY = 0$ if the child is female.

a. Write down the econometric specification of the linear regression model explaining *MATHSCORE* as a function of *SMALL*, *TCHEXPER*, *BOY* and $BOY \times TCHEXPER$, with parameters β_1, β_2, ...

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the *change* in the expected math score for a boy in a small class with a teacher having 11 years of experience rather than 10?

iv. What is the *change* in the expected math score for a boy in a small class with a teacher having 13 years of experience rather than 12?

v. State, in terms of the model parameters, the null hypothesis that the marginal effect of teacher experience on expected math score does not differ between boys and girls, against the alternative that boys benefit more from additional teacher experience. What test statistic would you use to carry out this test? What is the distribution of the test statistic assuming then null hypothesis is true, if $N = 1200$? What is the rejection region for a 5% test?

b. Modify the model in part (a) to include $SMALL \times BOY$.

i. What is the expected math score for a boy in a small class with a teacher having 10 years of experience?

ii. What is the expected math score for a girl in a regular-sized class with a teacher having 10 years of experience?

iii. What is the expected math score for a boy? What is it for a girl?

iv. State, in terms of the part (b) model parameters, the null hypothesis that the expected math score does not differ between boys and girls, against the alternative that there is a difference in expected math score for boys and girls. What test statistic would you use to carry out this test? What is the distribution of the test statistic assuming the null hypothesis is true, if $N = 1200$? What is the rejection region for a 5% test?

7.7 Can monetary policy reduce the impact of a severe recession? A **natural experiment** is provided by the State of Mississippi. In December of 1930, there were a series of bank failures in the southern United States. The central portion of Mississippi falls into two Federal Reserve Districts: the sixth (Atlanta Fed) and the eighth (St. Louis Fed). The Atlanta Fed offered "easy money" to banks while the St. Louis Fed did not. On July 1, 1930 (just before the crisis), there were 105 State Charter banks in Mississippi in the sixth district and 154 banks in the eighth district. On July 1, 1931 (just after the crisis), there were 96 banks remaining in the sixth district and 126 in the eighth district. These data values are from Table 1, Gary Richardson and William Troost (2009) "Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933," *Journal of Political Economy*, 117(6), 1031–1073.

a. Let the eighth district be the control group and the sixth district be the treatment group. Construct a figure similar to Figure 7.3 using the four observations rather than sample means. Identify the treatment effect on the figure.

b. How many banks did each district lose during the crisis? Calculate the magnitude of the treatment effect using (7.18) with these four observations, rather than sample means.

c. Suppose we have data on these two districts for 1929–1934, so $N = 12$. Let $AFTER_t = 1$ for years after 1930, and let $AFTER_t = 0$ for years 1929 and 1930. Let $TREAT_i = 1$ for the sixth district and let $TREAT_i = 0$ for banks in the eighth district. Let $BANKS_{it}$ be the number of banks in each district in each year. Angrist and Pischke (2015, p. 188) report the estimated equation

$$\underset{\text{(se)}}{\widehat{BANKS_{it}} = 167 - 2.9TREAT_i - 49AFTER_t + 20.5(TREAT_i \times AFTER_t)}$$
$$\text{(se)} \qquad (8.8) \qquad (7.6) \qquad (10.7)$$

Compare the estimated treatment effect from this equation to the calculation in (b). Is the estimated treatment effect significant, at the 5% level?

7.8 Using $N = 2005$ observations, we examine the relationship between food expenditures away from home per person in the past month as a function of household monthly income, the highest level of education of a household member, and region of the country. The full equation of interest is

$$\ln(FOODAWAY) = \beta_1 + \beta_2 \ln(INCOME) + \delta_1 COLLEGE + \delta_2 ADVANCED + \theta_1 MIDWEST + \theta_2 SOUTH + \theta_3 WEST + e$$

where $COLLEGE = 1$ if the highest education of a household member is a college degree, $ADVANCED = 1$ if the highest education of a household member is an advanced degree (such as a Master's or Ph.D.). The regional indicators equal one if the household lives in that region and are zero otherwise.

a. The estimated value of β_2 is 0.427 with a standard error of 0.035. Construct and interpret a 95% interval estimate.

b. The estimated value of δ_2 is 0.270 with a standard error of 0.0544. Construct and interpret a 95% interval estimate using the rough calculation in Section 7.3.1.

c. Use the exact calculation discussed in Section 7.3.1 to estimate the predicted effect on food expenditure per person away from home for a household having a member with an advanced degree.

d. What is the null hypothesis, in terms of the model parameters, that the highest level of education achieved by a household member does not matter? What is the test statistic for this hypothesis? What is the 5% rejection region? The sum of squared residuals from the full model is 1586 and *SSE* from the model omitting the education variables is 1609. Can we conclude that the education variables are important predictors of food expenditures away from home?

e. In the full model, the reported *t*-value for *COLLEGE* is 0.34. What can we conclude from that? [*Hint*: What is the reference group?]

f. The estimated value of θ_2 is 0.088. What is the estimated expected value of ln(*FOODAWAY*) for a household with $10,000 per month income, with a member with an advanced degree, and who live in the south? Calculate the natural and corrected predictors of expenditure on food away from home per member for this household. [*Hint*: A relevant piece of information is in part (b).]

7.9 Suppose we wish to estimate a model of household expenditures on alcohol (*ALC*, in dollars per month) as a function of household income (*INCOME*, $100's per month), and some other demographic variables.

a. Let *KIDS* = 0, 1, 2, … be the number children in the household. Is *KIDS* a qualitative or quantitative variable? Interpret the coefficient of *KIDS* in the model

$$ALC = \beta_1 + \beta_2 INCOME + \delta KIDS + e \tag{XR7.9.1}$$

What is the marginal impact of the second child? What is the marginal impact of the fourth child?

b. Let *ONEKID* = 1 if there is one child, and zero otherwise. Let *TWOKIDS* = 1 if there are two children, and zero otherwise. Let *MANY* = 1 if there are three or more children, and zero otherwise. Consider the model

$$ALC = \beta_1 + \beta_2 INCOME + \delta_1 ONEKID + \delta_2 TWOKIDS + \delta_3 MANY + e \tag{XR7.9.2}$$

Compare the interpretation of this model to that in part (a). Is the impact of an additional child the same as in the model in (a)? What is the impact of the first child on expected household expenditure on alcohol? What is the impact of having a fourth child on the expected household expenditure on alcohol?

c. Is there a set of parameter restrictions, or constraints, that we can impose on equation (XR7.9.2) to make it equivalent to equation (XR7.9.1)?

7.10 Suppose we wish to estimate a model of household expenditures on alcohol (*ALC*, in dollars per month) as a function of household income (*INCOME*, $100's per month), and some other demographic variables.

a. Let *RELIGIOUS* = 0, 1, 2, 3, or 4 if the household considers itself not religious, a little religious, moderately religious, very religious, or extremely religious, respectively. Is *RELIGIOUS* a quantitative or qualitative variable? Explain your choice.

b. Consider the model

$$ALC = \beta_1 + \beta_2 INCOME + \beta_3 RELIGIOUS + e$$

What is the expected household expenditure on alcohol for a household that considers itself not religious? What is the expected household expenditure for a household that considers itself a little religious? What is the expected household expenditure for a household that considers itself moderately religious?

c. If we test the hypothesis $\beta_3 = 0$ in model (b), what behavioral assumption are we testing? What is the expected household expenditure on alcohol if the hypothesis is true?

d. Let *LITTLE* = 1 if the household considers itself a little religious, and zero otherwise. Similarly define the indicator variables *MODERATELY*, *VERY*, and *EXTREMELY*. Consider the model

$$ALC = \gamma_0 + \gamma_1 INCOME + \gamma_2 LITTLE + \gamma_3 MODERATELY + \gamma_4 VERY + \gamma_5 EXTREMELY + e$$

What is the expected household expenditure for a household that considers itself not religious? What is the expected household expenditure for a household that considers itself a little religious? What is the expected household expenditure for a household that considers itself moderately religious? Very religious? Extremely religious?

e. If we impose the restrictions $\gamma_3 = 2\gamma_2$, $\gamma_4 = 3\gamma_2$, $\gamma_5 = 4\gamma_2$ on the model in part (d), how does the restricted model compare to the model in (b)?

7.11 Consider the log-linear regression model $\ln(y) = \beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D) + e$. If the regression errors are normally distributed $N\left(0, \sigma^2\right)$, then

$$E(y|x, D) = \exp\left(\beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D)\right) \exp\left(\sigma^2/2\right) \tag{XR7.11.1}$$

a. Use Derivative Rule 7 to show that

$$\frac{\partial E(y|x,D)}{\partial x} = \exp\left(\beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D)\right) \exp\left(\sigma^2/2\right)\left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.2)}$$

b. Divide both sides of the result in (a) by $E(y|x,D)$ to show that

$$\frac{\partial E(y|x,D)}{\partial x}\frac{1}{E(y|x,D)} = \frac{\partial E(y|x,D)/E(y|x,D)}{\partial x} = \left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.3)}$$

c. Multiply both sides of the equation in (b) by 100 to obtain

$$100\frac{\partial E(y|x,D)/E(y|x,D)}{\partial x} = \%\Delta E(y|x,D) = 100\left(\beta_2 + \delta_2 D\right) \quad \text{(XR7.11.4)}$$

This is the marginal effect, the percentage change, in $E(y|x,D)$ given a unit change in x in the log-linear model.

d. A fitted log-linear model for house price, where $SQFT(x)$ is the house's living area (100s of square feet) and $UTOWN(D)$ is an indicator variable with $UTOWN = 1$ for houses near a university, and zero otherwise, is

$$\widehat{\ln(PRICE)} = 4.456 + 0.362SQFT + 0.336UTOWN - 0.00349(SQFT \times UTOWN)$$

Use equation (XR7.11.4) to calculate the marginal effect of $SQFT$ on house price, for a house with $UTOWN = 1$ and for a house with $UTOWN = 0$.

e. Let b_2 and d_2 be the least squares estimators of β_2 and δ_2 in equation (XR7.11.4). Write down the formula for the standard error of the estimated value $100\left(b_2 + d_2 D\right)$, for a given D.

f. Multiply both sides in (XR7.11.3) by x, and by 100/100, and rearrange to obtain

$$\frac{\partial E(y|x,D)/E(y|x,D)}{\partial x}x = \frac{100\partial E(y|x,D)/E(y|x,D)}{100\partial x/x} = \left(\beta_2 + \delta_2 D\right)x \quad \text{(XR7.11.5)}$$

Interpreting $100\partial x/x$ as the percentage change in x, we find that the elasticity of expected price with respect to a percentage change in x is $\left(\beta_2 + \delta_2 D\right)x$.

g. Apply the result in equation (XR7.11.5) to calculate the elasticities of expected house price with respect to a change in price for a house of 2500 square feet, when $UTOWN = 1$ and when $UTOWN = 0$.

h. Let b_2 and d_2 be the least squares estimators of β_2 and δ_2 in equation (XR7.11.5). Write down the formula for the standard error of the estimated value $\left(b_2 + d_2 D\right)x$, given D and x.

7.12 Consider the log-linear regression model $\ln(y) = \beta_1 + \beta_2 x + \delta_1 D + \delta_2 (x \times D) + e$. If the regression errors are normally distributed $N\left(0, \sigma^2\right)$, then $E(y|x,D)$ is given in equation (XR7.11.1).

a. Find $E(y|x, D = 1)$ and $E(y|x, D = 0)$.

b. Show that

$$\frac{100\left[E(y|x,D=1) - E(y|x,D=0)\right]}{E(y|x,D=0)} = 100\left[\exp\left(\delta_1 + \delta_2 x\right) - 1\right] \quad \text{(XR7.12.1)}$$

This is the percentage change in the expected value of y, given x, when the indicator variable changes from $D = 0$ to $D = 1$.

c. Given the log-linear model, the value of $\ln(y)$ when $D = 0$ is $\ln(y|D = 0, x) = \beta_1 + \beta_2 x + e$, and when $D = 1$ we have $\ln(y|D = 1, x) = \left(\beta_1 + \delta_1\right) + \left(\beta_2 + \delta_2\right)x + e$. Subtract $\ln(y|D = 0, x)$ from $\ln(y|D = 1, x)$, and multiply by 100, to obtain

$$100\left[\ln(y|D=1,x) - \ln(y|D=0,x)\right] \simeq \%\Delta(y|x) = 100\left(\delta_1 + \delta_2 x\right) \quad \text{(XR7.12.2)}$$

d. A fitted log-linear model for house price, where $SQFT(x)$ is the house's living area (100s of square feet) and $UTOWN(D)$ is an indicator variable with $UTOWN = 1$ for houses near a university, and zero otherwise, is

$$\widehat{\ln(PRICE)} = 4.456 + 0.362SQFT + 0.336UTOWN - 0.00349(SQFT \times UTOWN)$$

Calculate the percentage change in the expected value of $PRICE$ for a house of 2500 square feet using (XR7.12.1). Also calculate the approximate value in (XR7.12.2).

e. If d_1 and d_2 are the least squares estimators of δ_1 and δ_2 in equation (XR7.12.2), write down the formula for the standard error of $100(d_1 + d_2 x)$, given x.

f. Let $\lambda = 100\left[\exp(\delta_1 + \delta_2 x) - 1\right]$ and $\hat{\lambda} = 100\left[\exp\left(d_1 + d_2 x\right) - 1\right]$. Use Derivative Rule 7, in Appendix A.3.1, to show that $\partial\lambda/\partial\delta_1 = 100\exp\left(\delta_1 + \delta_2 x\right)$ and $\partial\lambda/\partial\delta_2 = 100\exp(\delta_1 + \delta_2 x)x$. The "delta method" for finding the variance of a nonlinear function, such as $\hat{\lambda}$, is discussed in Section 5.7.4 and also Appendix 5B.5. Using the delta method, write out the expression for standard error of $\hat{\lambda}$.

7.13 Many cities in California have passed Inclusionary Zoning policies (also known as below-market housing mandates) as an attempt to make housing more affordable. These policies require developers to sell some units below the market price on a percentage of the new homes built. For example, in a development of 10 new homes each with market value $850,000, the developer may have to sell 5 of the units at $180,000. Means and Stringham (2012)[30] examine the effects of such policies on house prices and number of housing units available using 1990 and 2000 census data on 311 California cities.

a. Let *LNPRICE* be the log of average home price, and let *LNUNITS* be the log of the number of housing units. Using only the data for 2000, we compare the sample means of *LNPRICE* and *LNUNITS* for cities with an Inclusionary Zoning policy, *IZLAW* = 1, to those without the policy, *IZLAW* = 0. The following table displays the sample means of *LNPRICE* and *LNUNITS*.

2000	*IZLAW* = 1	*IZLAW* = 0
$\overline{LNPRICE}$	12.8914	12.2851
$\overline{LNUNITS}$	9.9950	9.5449

Based on these estimates, what is the percentage difference in prices and number of units for cities with and without the law? Use the approximation $100\left[\ln(y_1) - \ln\left(y_0\right)\right]$ for the percentage difference between y_0 and y_1. Does the law appear to achieve its purpose?

b. Using the data for 1990, we compare the sample means of *LNPRICE* and *LNUNITS* for cities with an Inclusionary Zoning policy, *IZLAW* = 1, to those without the policy, *IZLAW* = 0. The following table displays the sample means of *LNPRICE* and *LNUNITS*.

1990	*IZLAW* = 1	*IZLAW* = 0
$\overline{LNPRICE}$	12.3383	12.0646
$\overline{LNUNITS}$	9.8992	9.4176

Use the existence of an Inclusionary Zoning policy as a "treatment." Consider those cities that did not pass such a law, *IZLAW* = 0, the "control" group. Draw a figure similar to Figure 7.3 comparing treatment and control groups for *LNPRICE*, and determine the "treatment effect." Are your conclusions about the effect of the policy the same as in (a)?

c. Draw a figure similar to Figure 7.3 comparing treatment and control groups for *LNUNITS*, and determine the "treatment effect." Are your conclusions about the effect of the policy the same as in (a)?

7.14 Consider a model explaining the weekly sales (*SALES* = 100's cans sold) of a popular brand (the "target" brand) of canned tuna as a function of its price (*PRICE* = average price in cents), the average prices of two competitors (*PRICE*2, *PRICE*3, also in cents). Also included is an indicator variable *DISP* = 1 if there is a store display but **no** newspaper ad during the week for the target brand, and 0 otherwise. The indicator variable *DISPAD* = 1 if there is a store display during the week for the target

[30] Tom Means and Edward P. Stringham (2012) "Unintended or Intended consequences? The effect of below-market housing mandates on housing markets in California," *Journal of Public Finance and Public Choice*, p. 39–64. The authors wish to thank Tom Means for providing the data and insights into this exercise.

brand **and** newspaper ads, 0 otherwise. The estimated log-linear model is

$$\begin{array}{lllll}\widehat{\ln(SALES)} = 2.077 - 0.0375PRICE + 0.0115PRICE2 + 0.0129PRICE3 + 0.424DISP \\ \text{(se)} \quad (0.646) \quad (0.00577) \quad (0.00449) \quad (0.00605) \quad (0.105) \\ + 1.431DISPAD \quad R^2 = 0.84 \quad N = 52 \\ (0.156)\end{array}$$

a. Discuss and interpret the coefficients of the price variables.

b. Are the signs and relative magnitudes of the advertising variables consistent with economic logic? Provide both the "rough" and "exact" calculations for the effects of *DISP* and *DISPAD* from Sections 7.3.1 and 7.3.2.

c. Test the significance of the advertising variables using a two-tail test, at the 1% level of significance. What do you conclude?

d. The F-test statistic value for the joint significance of the two advertising variables is 42.0. What can we conclude about the significance of advertising? If you were going to use the form of the F-statistic in equation (6.4), what additional regression would you need to run?

e. Label the parameters in the equation $\beta_1, \beta_2, \ldots$ If the null hypothesis is $H_0: \beta_6 \leq \beta_5$, state the alternative hypothesis. Why is the test of this null hypothesis and alternative hypothesis interesting? Carry out the test at the 1% level of significance, given that the calculated t-value is 6.86. What do you conclude?

7.15 Mortgage lenders are interested in determining borrower and loan characteristics that may lead to delinquency or foreclosure. We estimate a regression model using 1000 observations and the following variables. The dependent variable of interest is *MISSED*, an indicator variable = 1 if the borrower missed at least three payments (90+ days late), but 0 otherwise. Explanatory variables are *RATE* = initial interest rate of the mortgage; *AMOUNT* = dollar value of mortgage (in \$100,000); and *ARM* = 1 if mortgage has an adjustable rate, and = 0 if mortgage has a fixed rate. The estimated equation is

$$\begin{array}{llll}\widehat{MISSED} = -0.348 + 0.0452RATE + 0.0732AMOUNT + 0.0834ARM \\ \text{(se)} \quad (0.00841) \quad (0.0144) \quad (0.0326)\end{array}$$

a. Interpret the signs and significance of each of the coefficients.

b. Two borrowers who did not miss a payment had loans with the following characteristics: ($RATE = 8.2$, $AMOUNT = 1.912$, $ARM = 1$) and ($RATE = 9.1$, $AMOUNT = 8.6665$, $ARM = 1$). For each of these borrowers, predict the probability that they will miss a payment.

c. Two borrowers who did miss a payment had loans with the following characteristics: ($RATE = 12.0$, $AMOUNT = 0.71$, $ARM = 0$) and ($RATE = 6.45$, $AMOUNT = 8.5$, $ARM = 1$). For each of these borrowers, predict the probability that they will miss a payment.

d. For a borrower seeking an adjustable rate mortgage, with an initial interest rate of 6.0, above what loan amount would you predict a missed payment with probability 0.51?

7.7.2 Computer Exercises

7.16 In this exercise, we examine the hours of market work by married women as a function of their education and number of children. Use data file *cps5mw_small* for this exercise. The data file *cps5mw* contains more observations.

a. Estimate the linear regression model

$$HRSWORK = \beta_1 + \beta_2 WAGE + \beta_3 EDUC + \beta_4 NCHILD + e \qquad \text{(XR7.16.1)}$$

Interpret the coefficient of *NCHILD*. Estimate the expected hours worked by a married woman whose wage is \$20 per hour, who has 16 years of education, and who has no children. Do the same calculation for a woman with one child, two children, and three children. How much does the expected number of hours change with each additional child?

b. Define the indicator variables $POSTGRAD = 1$ if $EDUC > 16, 0$ otherwise; $COLLEGE = 1$ if $EDUC = 16$, 0 otherwise; and *SOMECOLLEGE* if $12 < EDUC < 16$. Estimate the *HRSWORK* equation (XR7.16.1) replacing *EDUC* by these three indicator variables. Interpret the coefficients of the education indicator variables. Estimate the expected hours worked by a married woman

whose wage is \$20 per hour, who has 12 years of education, and who has no children. Do the same calculation for a woman with $EDUC = 13$, 14, 15, 16, and 17. Is the marginal effect of education constant?

c. Define indicator variables $ONEKID = 1$ if $NCHILD = 1$, 0 otherwise; $TWOKIDS = 1$ if $NCHILD = 2$, 0 otherwise; and $MOREKIDS = 1$ if $NCHILD > 2$, 0 otherwise. Estimate the *HRSWORK* equation (XR7.16.1) but replace *NCHILD* by these three indicator variables. Interpret the estimated coefficients of the three indicator variables. Estimate the expected hours worked by a married woman with 16 years of education, whose wage is \$20 per hour with no children, one child, two children, and more than two children. Compare and contrast these estimates to those in (a).

d. Estimate the model (XR7.16.1) replacing *EDUC* with the three indicator variables in (b) and replacing *NCHILD* with the three indicator variables in (c). Compare and contrast this model to the models in (a)–(c).

e. Define the indicator variable $EDUC12 = 1$ if $EDUC = 12$, 0 otherwise. Define indicator variables *EDUC12*, *EDUC13*, *EDUC14*, *EDUC16* similarly. In this sample, there are no women with 15 years of education. Define $EDUC18 = 1$ if $EDUC > 16$, 0 otherwise. Estimate the *HRSWORK* equation (XR7.16.1) replacing *NCHILD* by the three indicator variables and *EDUC* by the five new indicator variables. Have any essential conclusions changed by using this specification?

f. Which of the specifications in (a)–(e) has the highest R^2? The highest adjusted-R^2, the smallest *SCHWARZ* criterion (SC or BIC) value? Which model do you prefer taking into account economic, econometric, and fit aspects?

7.17 Does a mother's smoking affect the birthweight of her child? Using data in the file *bweight_small* taken from Cattaneo (2010),[31] we explore this question. The file *bweight* contains more observations.

a. Calculate the sample means of *BWEIGHT* for mothers who smoke ($MBSMOKE = 1$) and those who do not smoke ($MBSMOKE = 0$). Use the t-test of the equality of population means given in Appendix C.7.2, Case 1, to test whether the mean birthweight for smoking and nonsmoking mothers is the same. Use the 5% level of significance.

b. Estimate the regression $BWEIGHT = \beta_1 + \beta_2 MBSMOKE + e$. Interpret the coefficient of *MBSMOKE*. Can we interpret the coefficient as the "average treatment effect" of smoking? Test the null hypothesis that $\beta_2 \geq 0$ against $\beta_2 < 0$ at the 5% level of significance.

c. Add to the model in (b) control variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Are any of these variables significant predictors of an infant's birthweight? Which signs of the significant coefficients are consistent with your expectations? Does the estimate of the coefficient of *MBSMOKE* change much?

d. Estimate the regression of *BWEIGHT* on *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY* for mothers who smoke ($MBSMOKE = 1$) and those who do not smoke ($MBSMOKE = 0$). Carry out a Chow test of the equivalence of these two regressions at the 5% level.

e. Use equation (7.37) to obtain the estimate of the average treatment effect using the results from (d). Compare this estimate of the average treatment effect to the estimates in (b) and (c).

7.18 Does a mother's smoking affect the birthweight of her child? Using the data file *bweight_small*, we explore this question. The file *bweight* contains more observations.

a. Estimate the regression model represented by equation (7.38) for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MBSMOKE* and interactions between *MBSMOKE* and the other variables. Use equation (7.40), and the discussion below equation (7.40), to estimate the average treatment effect.

b. Use equation (7.41) to estimate the average treatment effect of mother smoking on infant birthweight, and construct a 95% interval estimate for τ_{ATE}.

c. Calculate the normalized difference equation (7.44) for each of the variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Are any of the normalized differences bigger than the rule of thumb threshold of 0.25?

d. Use equation (7.42) to estimate the average treatment effect on the treated, τ_{ATT}. How much does it differ from your estimate of the population average treatment effect?

[31] Efficient semiparametric estimation of multi-valued treatment effects under ignorability, *Journal of Econometrics*, 155, 138–154. The authors would like to thank Matias Cattaneo for providing the data. The dataset is used in *Stata Treatment-Effects Reference Manual, Release 14* for examples as well.

e. Use equation (7.43) to estimate the average treatment effect on the population of mothers who are Hispanic (*MHISP* = 1). How does it compare to the estimated population average treatment effect?

f. Use equation (7.43) to estimate the average treatment effect on the population of mothers who are white (*MWHITE* = 1). How does this compare to the population average treatment effect estimate?

7.19 Does a mother's smoking affect the birthweight of her child? Using the data file *bweight_small* we explore this question. The file *bweight* contains more observations. The variable *MSMOKE* is the number of cigarettes smoked daily during pregnancy. Nonsmokers (*MBSMOKE* = 0) smoke zero daily. Among smokers (*MBSMOKE* = 1), the variable *MSMOKE* = 1 if 1–5 cigarettes are smoked daily; *MSMOKE* = 2 if 6–10 cigarettes are smoked daily; and *MSMOKE* = 3 if 11 or more cigarettes are smoked daily.

a. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MSMOKE*. Interpret the estimated coefficient of *MSMOKE*.

b. From *MSMOKE* create three indicator variables, *SMOKE2* = 1 if a mother smokes 1–5 cigarettes per day, 0 otherwise; *SMOKE3* = 1 if a mother smokes 6–10 cigarettes per day, 0 otherwise; *SMOKE4* = 1 if a mother smokes 11 or more cigarettes per day, 0 otherwise. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *SMOKE2*, *SMOKE3*, and *SMOKE4*. Interpret the estimated coefficients of *SMOKE2*, *SMOKE3*, and *SMOKE4*. Does smoking 1–5 cigarettes per day have a statistically significant negative effect on infant birthweight?

c. Using the results in (b), test the null hypothesis that smoking 11 or more cigarettes per day reduces birthweight by no more than smoking 6–10 cigarettes per day, against the alternative that smoking 11 or more cigarettes per day reduces birthweight by more than smoking 6–10 cigarettes per day.

d. Using the results in (b), test the null hypothesis that smoking 11 or more cigarettes per day reduces birthweight by no more than smoking 1–5 cigarettes per day, against the alternative that smoking 11 or more cigarettes per day reduces birthweight by more than smoking 1–5 cigarettes per day.

e. Estimate a regression model for *BWEIGHT*. Include as explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*. Estimate the model separately for *MSMOKE* = 0, 1, 2, and 3. Using each model, estimate the expected birthweight of a child of a married woman who is 25 years old whose first prenatal visit was in the first trimester and who had already given birth to at least one child. What do you observe?

f. Estimate the linear probability model with dependent variable *LBWEIGHT* as a function of explanatory variables *MMARRIED*, *MAGE*, *PRENATAL1*, and *FBABY*, along with *MSMOKE*. Predict the probability of a low-birthweight infant for *MSMOKE* = 0, 1, 2, and 3 of a married woman who is 25 years old whose first prenatal visit was in the first trimester and who had already given birth to at least one child. What do you observe?

7.20 In this exercise, we will explore some of the factors predicting costs at American universities using the data file *poolcoll2* and observations outside the great recession. Let *TC* = the real ($2008) total cost per student, *FTUG* = number of full-time undergraduate students, *FTGRAD* = number of full-time graduate students, *FTEF* = full-time faculty per 100 students, *CF* = number of contract faculty per 100 students, *FTENAP* = full-time nonacademic professionals per 100 students.

a. Estimate the regression of ln(*TC*) on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?

b. What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?

c. Add the indicator variable *PRIVATE* to the model. Do you predict higher or lower total cost per student at private universities? Is this a statistically significant factor in predicting total cost per student?

d. Add to the model not only *PRIVATE* but also *PRIVATE* × *FTEF*. Are these variables individually and jointly significant at the 5% level?

e. Add to the model not only *PRIVATE* but also *PRIVATE* times all the other variables. Test the joint significance of *PRIVATE* and *PRIVATE* times all the other variables using an *F*-test. What do you conclude about the model in (a) that does not distinguish between private and public universities?

f. Estimate the model in (a) twice, once for private universities and once for public universities. Call the sum of squared residuals for the private universities *SSE*1, and the sum of squared residuals for the public universities *SSE*0. Compare *SSE*1 + *SSE*0 to the sum of squared residuals in part (e).

7.21 In this exercise, we explore some of the factors predicting costs at American public universities using the data file *pubcoll*. Let TC = the real ($2008) total cost per student, $FTUG$ = number of full-time undergraduate students, $FTGRAD$ = number of full-time graduate students, $FTEF$ = full-time faculty per 100 students, CF = number of contract faculty per 100 students, and $FTENAP$ = full-time nonacademic professionals per 100 students.

a. Estimate the regression of ln(TC) on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?

b. What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?

c. Add indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. Are these variables jointly and individually significant? Using your favorite site for macroeconomic data, plot the quarterly percentage change in the real U.S. GDP from January 1987 to January 1993. Does this help explain the signs and significance of any of the indicator variable coefficients?

d. The variable $CRASH = 1$ during 2008, 2010, and 2011. Add to the model in (c) interactions between *CRASH* and each of the variables *FTEF*, *CF*, and *FTENAP*. Are these variables individually significant at the 5% level? Are they jointly significant?

e. Add to the model in (d) interactions between *CRASH* and each of the variables *FTUG* and *FTGRAD*. Considering all the interaction variables, which are significant at the 5% level? Test the joint significance of all the interaction variables at the 5% level.

7.22 In this exercise, we explore some of the factors predicting costs at American public universities using the data file *pubcoll*. Let TC = the real ($2008) total cost per student, $FTUG$ = number of full-time undergraduate students, $FTGRAD$ = number of full-time graduate students, $FTEF$ = full-time faculty per 100 students, CF = number of contract faculty per 100 students, and $FTENAP$ = full-time nonacademic professionals per 100 students. Use only the data for years prior to 2008. Include in the model year indicator variables *D1989*, *D1991*, *D1999*, and *D2005*.

a. Estimate the regression of ln(TC) on the remaining variables. What are the predicted effects of additional undergraduate students and graduate students on total cost per student?

b. What are the predicted effects of additional full-time faculty, contract faculty, and nonacademic professionals on total cost per student?

c. Using the estimates from part (a), compute the normal and corrected predictors of total cost using 2005 data for University of Arizona (unitid = 104179), Indiana University-Bloomington (unitid 151351), and The University of Texas at Austin (unitid = 228778). Compare the predicted values to the reported TC for 2005. Which schools had actual total cost TC higher than predicted?

d. Add an indicator variable for each different university except the first, which is the reference group. Test the joint significance of these indicator variables at the 5% level of significance using the F-test given in equation (6.4). Are there individual differences among the universities?

e. Using the estimates from part (d), compute the normal and corrected predictors of total cost using 2005 data for University of Arizona (unitid = 104179), Indiana University-Bloomington (unitid 151351), and The University of Texas at Austin (unitid = 228778). Compare the predicted values to the reported TC for 2005. Which schools had actual total cost TC higher than predicted?

7.23 In the STAR experiment (Section 7.5.3), children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star5_small2*.

a. Calculate the average of *MATHSCORE* for (i) students in regular-sized classrooms with full-time teachers but no aide; (ii) students in regular-sized classrooms with full-time teachers and an aide; and (iii) students in small classrooms. What do you observe about test scores in these three types of learning environments?

b. Estimate the regression model $MATHSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + e_i$, where *AIDE* is an indicator variable equaling 1 for classes taught by a teacher and an aide, and 0 otherwise. What is the relation of the estimated coefficients from this regression to the sample means in part (a)? Test the statistical significance of β_3 at the 5% level.

c. To the regression in (b) add the additional explanatory variable *TCHEXPER*. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3? Construct a 95% interval estimate of expected math score for a student in a small class with a teacher having

10 years of experience. Construct a 95% interval estimate of expected math score for a student in a class with an aide and having a teacher with 10 years of experience. Calculate the least squares residuals from this model, calling them *EHAT*. This variable will be used in the next part.

d. To the regression in (c), add the additional indicator variable *FREELUNCH*. Students from lower income households receive a free lunch at school. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3? What explains the sign of *FREELUNCH*? Calculate the sample average of *EHAT*, from part (c), for students receiving a free lunch, and for students who do not receive a free lunch. Are the residual averages consistent with the regression that includes *FREELUNCH*?

e. To the model in (d), add interaction variables between *FREELUNCH* and *SMALL*, *AIDE* and *TCHEXPER*. Are any of these individually significant? Test the joint significance of these three interaction variables at the 5% level. What do you conclude?

f. Carry out a Chow test for the equivalence of the regression $MATHSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + \beta_4 TCHEXPER + e_i$ for students who receive a free lunch and those who do not receive a free lunch. How does this test result compare to the test result in part (e)?

7.24 Many cities in California have passed Inclusionary Zoning policies (also known as below-market housing mandates) as an attempt to make housing more affordable. These policies require developers to sell some units below the market price on a percentage of the new homes built. For example, in a development of 10 new homes each with market value $850,000, the developer may have to sell 5 of the units at $180,000. Means and Stringham (2012), and exercise 7.13, examine the effects of such policies on house prices and number of housing units available using 1990 and 2000 census data on California cities. Use the data file *means* for the following exercises.

a. Use *LNPRICE* and *LNUNITS* as dependent variables in difference-in-difference regressions, with explanatory variables *D*, the indicator variable for year 2000; *IZLAW*, and the interaction of *D* and *IZLAW*. Is the estimate of the treatment effect statistically significant, and of the anticipated sign?

b. To the regressions in (a) add the control variable *LMEDHHINC*. Interpret the estimate of the new variable, including its sign and significance. How does this addition affect the estimates of the treatment effect?

c. To the regressions in (b) add the variables 100(*EDUCATTAIN*), 100(*PROPPOVERTY*), and *LPOP*. Interpret the estimates of these new variables, including their signs and significance. How do these additions affect the estimates of the treatment effect?

d. Consider the differences-in-differences regression for *LNPRICE*

$$\ln(PRICE_{it}) = \beta_1 + \beta_2 IZLAW_i + \beta_3 D_t + \delta(IZLAW_i \times D_t) + \theta CITY_i + e_{it}$$

In this model, $CITY_i$ represents some unobservable characteristic of each city that stays constant over time. Write this model for the year 2000 $(D_t = 1)$. Write this model for the year 1990 $(D_t = 0)$. Subtract the expression for 1990 from the expression for 2000. The dependent variable is

$$DLNPRICE_i = \left[\ln(PRICE_{i,2000}) - \ln(PRICE_{i,2000})\right] \simeq \%\Delta PRICE_i/100$$

which is the decimal equivalent of the percentage change in price for city *i*. What parameters and variables remain on the right-hand side after the subtraction?

e. Regress $DLNPRICE_i$ against $IZLAW_i$ and compare the result to the *LNPRICE* regression in part (a).

7.25 Professor Ray C. Fair's voting model was introduced in Exercise 2.23. He builds models that explain and predict the U.S. presidential elections. See his website at http://fairmodel.econ.yale.edu/vote2016/index2.htm and see in particular his paper entitled "Presidential and Congressional Vote-Share Equations: November 2014 Update." The basic premise of the model is that the Democratic party's share of the two-party [Democratic and Republican] popular vote is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for reelection. Data for 1916–2016 are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic party. In addition to *GROWTH* and *INFLAT*, the explanatory variables include the following:

INCUMB = 1 if there is a Democratic incumbent at the time of the election and −1 if there is a Republican incumbent.

GOODNEWS = (number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% at an annual rate except for 1920, 1944, and 1948, where the values are zero) × *INCUMB*.

DPER = 1 if the incumbent is running for election and 0 otherwise.
DUR = 0 if the Democratic party has been in power for one term, 1[−1] if the Democratic [Republican] party has been in power for two consecutive terms, 1.25[−1.25] if the Democratic [Republican] party has been in power for three consecutive terms, 1.50 for four consecutive terms, and so on.
WAR = 1 for the elections of 1920, 1944, and 1948 and 0 otherwise.

a. Consider the regression model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLAT + \beta_4 GOODNEWS + \beta_5 DPER + \beta_6 DUR + \beta_7 INCUMB + \beta_8 WAR + e$$

Discuss the anticipated effects of the dummy variable *DPER*.

b. The variable *INCUMB* is somewhat different than dummy variables we have considered. Write out the regression function $E(VOTE)$ when there is a Democratic incumbent. Write out the regression function $E(VOTE)$ when there is a Republican incumbent. Recall that the signs of *GOODNEWS*, *GROWTH*, and *INFLAT* depend on *INCUMB*. Discuss the effects of this specification.

c. Use the data for the period 1916–2012 to estimate the proposed model. Discuss the estimation results. Are the signs as expected? Are the estimates statistically significant? How well does the model fit the data?

d. Use the regression result from part (c) to predict the value of *VOTE* for the 2016 election using the actual values of the explanatory variables.

e. Use the regression result from part (c) to construct a 95% prediction interval for the value of *VOTE* for the 2016 election using the actual values of the explanatory variables.

f. Use the data for the period 1916–2012 to estimate the proposed model. In election year 2016, *INCUMB* = 1, *DPER* = 0, *DUR* = 1, and *WAR* = 0. Using *GROWTH* = 2.16, *INFLAT* = 1.37, and *GOODNEWS* = 3, predict the vote in favor of the Democratic party candidate in 2016.

g. Using the results in (f), predict the vote in favor of the Democratic party in 2016 if *GOODNEWS* = 3, *GROWTH* = 2.16, and *INFLAT* = 0.

h. Using the results in (f), predict the vote in favor of the Democratic party in 2016 if *GOODNEWS* = 3, *GROWTH* = 4.0, and *INFLAT* = 0.

7.26 The data file *br2* contains data on 1080 house sales in Baton Rouge, Louisiana, during July and August 2005. The variables are: *PRICE* ($), *SQFT* (total square feet), *BEDROOMS* (number), *BATHS* (number), *AGE* (years), *OWNER* (= 1 if occupied by owner; 0 if vacant or rented), *TRADITIONAL* (= 1 if traditional style; 0 if other style), *FIREPLACE* (= 1 if present), *WATERFRONT* (= 1 if on waterfront).

a. Compute the data summary statistics and comment. In particular, construct a histogram of *PRICE*. What do you observe?

b. Estimate a regression model explaining ln(*PRICE*/1000) as a function of the remaining variables. Divide the variable *SQFT* by 100 prior to estimation. Comment on how well the model fits the data. Discuss the signs and statistical significance of the estimated coefficients. Are the signs what you expect? Give an exact interpretation of the coefficient of *WATERFRONT*.

c. Create a variable that is the product of *WATERFRONT* and *TRADITIONAL*. Add this variable to the model and reestimate. What is the effect of adding this variable? Interpret the coefficient of this interaction variable and discuss its sign and statistical significance.

d. It is arguable that the traditional style homes may have a different regression function from the diverse set of nontraditional styles. Carry out a Chow test of the equivalence of the regression models for traditional versus nontraditional styles. What do you conclude?

e. Predict the value of a traditional style house with 2500 square feet of area, that is 20 years old, which is owner occupied at the time of sale, with a fireplace, but no pool, and not on the waterfront.

7.27 The three most important words in real estate are "location, location, location!" We explore this question using 500, single-family home sales in Baton Rouge, LA from 2009 to 2013 in the data file *collegetown*. See *collegetown.def* for variable definitions.

a. Estimate the log-log model $\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_1 CLOSE + e$. Interpret the estimated coefficients of $\ln(SQFT)$ and *CLOSE*. Is the location variable *CLOSE* statistically significant at the 5% level?

b. Estimate the log-log model $\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_2 [CLOSE \times \ln(SQFT)] + e$. Interpret the estimated coefficients of $\ln(SQFT)$ and $[CLOSE \times \ln(SQFT)]$. Is the location variable $[CLOSE \times \ln(SQFT)]$ statistically significant at the 5% level?

c. Estimate the log-log model

$$\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \delta_1 CLOSE + \delta_2 [CLOSE \times \ln(SQFT)] + e$$

Are the location variables *CLOSE* and $[CLOSE \times \ln(SQFT)]$ individually and jointly statistically significant at the 5% level?

d. Using the model in (c), predict the prices of two houses with 2500 square feet, one close to the university and another that is not close. Use the corrected predictor.

e. Add *FIREPLACE*, *TWOSTORY*, and *OCCUPIED* to the model in (c). How do these features affect the price of a house?

f. Carry out a Chow test for the log-log model, comparing houses that are close to the university to those that are not close, using explanatory variables ln(*SQFT*), *FIREPLACE*, *TWOSTORY*, and *OCCUPIED*. What is the *p*-value of the test?

7.28 How much of an incumbency advantage do winners in U.S. House elections enjoy? This is the topic of a paper by David S. Lee (2008) "Randomized experiments from nonrandom selection in U.S. House elections," *Journal of Econometrics*, 142(2), 675–697. Lee uses a regression discontinuity approach to estimate the effect. There are 435 Congressional districts in the United States and elections are held every 2 years. Representatives serve a term of 2 years. We employ a subset of Lee's data. The data file *rddhouse_small* has 1200 observations. See the *rddhouse_small.def* for data details. The data file *rddhouse* is larger. The forcing variable is *SHARE*, which is the Democratic share of the votes in a election in year t minus 0.50, so that *SHARE* is the Democratic margin of victory. The outcome of interest is the Democratic share of the vote in the next election, *SHARENEXT*.

a. Create a scatter plot with *SHARE* on the horizontal axis and *SHARENEXT* on the vertical axis. Does there appear to be positive relationship, an inverse relationship, or no relationship?

b. The dummy variable $D = 1$ if $SHARE > 0$ and $D = 0$ if $SHARE < 0$. Estimate the regression model with *SHARENEXT* as dependent variable, and *SHARE*, *D,* and $SHARE \times D$ as explanatory variables. Interpret the magnitudes, signs, and significance of the coefficients of *D* and $SHARE \times D$. Graph the fitted value from this regression against *SHARE*.

c. The variable *BIN* is the center of an interval of width 0.005, starting at −0.25. There are 100 bins between −0.25 and 0.25. Define a "narrow" win or loss as being an election where the margin of victory, or loss, is within the interval −0.005 to 0.005. Calculate the sample means of *SHARENEXT* when $BIN = -0.0025$ and when $BIN = 0.0025$. Is the difference in means an estimate of the value of incumbency? Explain how.

d. Treat the two groups created in (c) as two populations. Carry out a test of the difference between the two population means using the test in Appendix C.7.2, Case 1. Using a two-tail test and the 5% level of significance, do we reject the equality of the two population means, or not?

e. The variables *SHARE2*, *SHARE3*, and *SHARE4* are *SHARE* raised to the second, third, and fourth power, respectively. Estimate the regression model with *SHARENEXT* as dependent variable, with explanatory variables *SHARE* and its powers, *D* and *D* times *SHARE* and its powers. Interpret the magnitudes, signs, and significance of the coefficients of *D*, and *D* times *SHARE*.

f. Graph the fitted value from the regression in (e) against *SHARE*. Is the fitted line similar to the one in (b)?

g. Estimate the regression with *SHARENEXT* as dependent variable with explanatory variables *SHARE* and its powers, for the observations when $D = 0$. Reestimate the regression for the observations when $D = 1$. Compare these results to those in (e).

h. The variable *BIN* in part (c) was created using the equation $BIN = SHARE - \text{mod}(SHARE, 0.005) + 0.0025$, where "mod" is the "modulus operator," a common software function. In particular, $\text{mod}(x, y) = x - y \times \text{floor}(x/y)$ where the operator "floor" rounds the argument down to the next integer. Explain how this operator works in this application to create "bins" of width 0.005.

7.29 How much of an incumbency advantage do winners in U.S. Senate elections enjoy? This issue is examined by Matias D. Cattaneo, Brigham R. Frandsen and Rocío Titiunik (2015) "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate, *Journal of Causal Inference*, 3(1): 1–24.[32] As they describe (p. 11): "Term length in the U.S. Senate is 6 years and there are 100 seats. These Senate seats are divided into three classes of

[32] Also in "Robust Data-Driven Inference in the Regression-Discontinuity Design," by Sebastian Calonico, Matias D. Cattaneo and Rocio Titiunik, *Stata Journal* 14(4): 909–946, 4th Quarter 2014.

roughly equal size (Class I, Class II, and Class III), and every 2 years only the seats in one class are up for election. As a result, the terms are staggered: In every general election, which occurs every 2 years, only one-third of Senate seats are up for election. Each state elects two senators in different classes to serve a 6-year term in popular statewide elections. Since its two senators belong to different classes, each state has Senate elections separated by alternating 2-year and 4-year intervals." We employ a subset of their data, contained in the file *rddsenate*. See *rddsenate.def* for data details. The forcing variable is *MARGIN*, which is the Democratic share of the votes in an election in year t minus 50: it is the Democratic margin of victory. The outcome of interest is the Democratic share of the vote in the next election for that Senate seat, *VOTE*.

a. Create a scatter plot with *MARGIN* on the horizontal axis and *VOTE* on the vertical axis. Does there appear to be a positive relationship, an inverse relationship, or no relationship?

b. The dummy variable $D = 1$ if $MARGIN > 0$ and $D = 0$ if $MARGIN < 0$. Estimate the regression model with *VOTE* as dependent variable, and *MARGIN*, D, and $MARGIN \times D$ as explanatory variables. Interpret the magnitudes, signs, and significance of the coefficients of D and $MARGIN \times D$. Graph the fitted value from this regression against *MARGIN*.

c. The variable *BIN* is the center of an interval of width 5, starting at -97.5 and ending at 102.5. Define a "narrow" win or loss as being an election where the margin of victory, or loss, is within the interval -2.5 to 2.5. Calculate the sample means of *VOTE* when $BIN = -2.5$ and when $BIN = 2.5$. Is the difference in means an estimate of the value of incumbency? Explain how.

d. Treat the two groups created in (c) as two populations. Carry out a test of the difference between the two population means using the test in Appendix C.7.2, Case 1: Using a two-tail test and the 5% level of significance, do we reject the equality of the two population means, or not?

e. The variables *MARGIN2*, *MARGIN3*, and *MARGIN4* are *MARGIN* raised to the second, third, and fourth powers, respectively. Estimate the regression model with *VOTE* as dependent variable, with explanatory variables *MARGIN* and its powers, D and D times *MARGIN* and its powers. Interpret the magnitudes, signs, and significance of the coefficients of D and D times *MARGIN*.

f. Graph the fitted value from the regression in (e) against *MARGIN*. Is the fitted line similar to the one in (b)?

g. How would the results of (e) compare to the regression with *VOTE* as dependent variable with explanatory variables *MARGIN* and its powers, for the observations when $D = 0$. What if the regression was estimated for the observations when $D = 1$?

7.30 What effect does having public health insurance have on the number of doctor visits a person has during a year? Using 1988 data, *rwm88_small*, from Germany we will explore this question. The data file *rwm88* contains more observations. The data were used by Regina T. Riphahn, Achim Wambach, and Andreas Million, "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, Vol. 18, No. 4, 2003, pp. 387–405.

a. Construct a histogram of *DOCVIS*. How many doctor visits do most patients in the survey have during the year? What are the mean and median number of doctor visits? What is the 90th percentile?

b. Test the null hypothesis that the population mean number of doctor visits for those with public insurance is the same as those who do not have public insurance. Use the 5% level of significance and a one-tail test.

c. Estimate the regression model with dependent variable *DOCVIS* and explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*. Comment on the signs and significance of these predictor variables.

d. Estimate the regression model with dependent variable *DOCVIS* and explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* separately for those with public insurance and those who do not have public insurance. Use equation (7.37) to obtain the estimate of the average treatment effect of public insurance.

e. Estimate the regression model with dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* in "deviation from the mean" form. That is, for each variable x create the variable $\tilde{x} = x - \bar{x}$, where $\bar{x}$ is the sample mean. Compare these results to those in (c).

f. Estimate the regression model with dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*, along with *PUBLIC* and *PUBLIC* times each of the variables in deviation about the mean form. What is the estimated average treatment effect? Is it statistically significant at the 5% level?

Appendix 7A Details of Log-Linear Model Interpretation

You may have noticed that in Section 7.3, while discussing the interpretation of the log-linear model, we omitted the error term, and we did not discuss the regression function $E(WAGE|\mathbf{x})$. To do so, we make use of the properties of the log-normal distribution in Appendix B.3.9 and discussed in Problem 7.11. There we noted that for the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$, if the error term $e \sim N(0, \sigma^2)$, then the expected value of y is

$$E(y|\mathbf{x}) = \exp(\beta_1 + \beta_2 x + \sigma^2/2) = \exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)$$

Starting from this equation, we can explore the interpretation of dummy variables and interaction terms.

Let D be a dummy variable. Adding this to our log-linear model, we have $\ln(y) = \beta_1 + \beta_2 x + \delta D + e$ and

$$E(y|\mathbf{x}) = \exp(\beta_1 + \beta_2 x + \delta D) \times \exp(\sigma^2/2)$$

If we let $E(y_1|\mathbf{x})$ and $E(y_0|\mathbf{x})$ denote the cases when $D = 1$ and $D = 0$, respectively, then we can compute their percentage difference as

$$\begin{aligned}
\%\Delta E(y|\mathbf{x}) &= 100\left[\frac{E(y_1|\mathbf{x}) - E(y_0|\mathbf{x})}{E(y_0|\mathbf{x})}\right]\%, \\
&= 100\left[\frac{\exp(\beta_1 + \beta_2 x + \delta) \times \exp(\sigma^2/2) - \exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}{\exp(\beta_1 + \beta_2 x) \times \exp(\sigma^2/2)}\right]\% \\
&= 100\left[\frac{\exp(\beta_1 + \beta_2 x)\exp(\delta) - \exp(\beta_1 + \beta_2 x)}{\exp(\beta_1 + \beta_2 x)}\right]\% = 100[\exp(\delta) - 1]\%
\end{aligned}$$

The interpretation of dummy variables in log-linear models carries over to the regression function. The percentage difference in the *expected* value of y is $100[\exp(\delta) - 1]\%$.

Appendix 7B Derivation of the Differences-in-Differences Estimator

To verify the expression for the differences-in-differences estimator in (7.14), note that the numerator can be expressed as

$$\begin{aligned}
\sum_{i=1}^{N}(d_i - \bar{d})(y_i - \bar{y}) &= \sum_{i=1}^{N} d_i(y_i - \bar{y}) - \bar{d}\sum_{i=1}^{N}(y_i - \bar{y}) \\
&= \sum_{i=1}^{N} d_i(y_i - \bar{y}) \quad \left[\text{using } \sum_{i=1}^{N}(y_i - \bar{y}) = 0\right] \\
&= \sum_{i=1}^{N} d_i y_i - \bar{y}\sum_{i=1}^{N} d_i \\
&= N_1\bar{y}_1 - N_1\bar{y} \\
&= N_1\bar{y}_1 - N_1(N_1\bar{y}_1 + N_0\bar{y}_0)/N \\
&= \frac{N_0 N_1}{N}(\bar{y}_1 - \bar{y}_0) \quad [\text{using } N = N_1 + N_0]
\end{aligned}$$

The denominator of b_2 is

$$\begin{aligned}\sum_{i=1}^{N}\left(d_i-\bar{d}\right)^2 &= \sum_{i=1}^{N}d_i^2-2\bar{d}\sum_{i=1}^{N}d_i+\sum_{i=1}^{N}\bar{d}^2\\ &= \sum_{i=1}^{N}d_i-2\bar{d}N_1+N\bar{d}^2 \quad \left[\text{using } d_i^2=d_i \text{ and } \sum_{i=1}^{N}d_i=N_1\right]\\ &= N_1-2\frac{N_1}{N}N_1+N\left(\frac{N_1}{N}\right)^2\\ &= \frac{N_0N_1}{N} \quad \left[\text{using } N=N_0+N_1\right]\end{aligned}$$

Combining the expressions for numerator and denominator, we obtain the result for the difference estimator in (7.14).

Appendix 7C The Overlap Assumption: Details

To see the impact of the difference of means, $\bar{x}_1-\bar{x}_0$, on the average treatment effect we begin with the separate regressions on the control and treatment groups used to compute the average treatment effect in Section 7.6.4, $\hat{\alpha}_0+\hat{\beta}_0x_i$ and $\hat{\alpha}_1+\hat{\beta}_1x_i$. Using the property of least squares fitted lines, the estimated intercepts are

$$\hat{\alpha}_0=\bar{y}_0-\hat{\beta}_0\bar{x}_0 \text{ and } \hat{\alpha}_1=\bar{y}_1-\hat{\beta}_0\bar{x}_1$$

We can express the sample mean of the control variable as

$$\begin{aligned}\bar{x}=N^{-1}\sum_{i=1}^{N}x_i &= N^{-1}\left[\sum_{i=1}^{N_0}x_i+\sum_{i=N_0+1}^{N}x_i\right]=N^{-1}\left[N_0\bar{x}_0+N_1\bar{x}_1\right]\\ &= \frac{N_0\bar{x}_0}{N}+\frac{N_1\bar{x}_1}{N}=f_0\bar{x}_0+f_1\bar{x}_1\end{aligned}$$

The control variable sample mean $\bar{x}$ is a weighted average of $\bar{x}_0$ and $\bar{x}_1$, where the weight f_0 is the fraction of the observations in the control group and f_1 is the fraction of observations in the treatment group. Then

$$\begin{aligned}\hat{\tau}_{ATE} &= \left(\hat{\alpha}_1-\hat{\alpha}_0\right)+\left(\hat{\beta}_1-\hat{\beta}_0\right)\bar{x}\\ &= \left[\left(\bar{y}_1-\hat{\beta}_1\bar{x}_1\right)-\left(\bar{y}_0-\hat{\beta}_0\bar{x}_0\right)\right]+\left(\hat{\beta}_1-\hat{\beta}_0\right)\left(f_0\bar{x}_0+f_1\bar{x}_1\right)\\ &= \left(\bar{y}_1-\bar{y}_0\right)-\hat{\beta}_1\bar{x}_1+\hat{\beta}_0\bar{x}_0+f_0\hat{\beta}_1\bar{x}_0+f_1\hat{\beta}_1\bar{x}_1-f_0\hat{\beta}_0\bar{x}_0-f_1\hat{\beta}_0\bar{x}_1\\ &= \left(\bar{y}_1-\bar{y}_0\right)+\left(f_1\hat{\beta}_1\bar{x}_1-\hat{\beta}_1\bar{x}_1\right)-\left(f_0\hat{\beta}_0\bar{x}_0-\hat{\beta}_0\bar{x}_0\right)+f_0\hat{\beta}_1\bar{x}_0-f_1\hat{\beta}_0\bar{x}_1\\ &= \left(\bar{y}_1-\bar{y}_0\right)+\left(f_1-1\right)\hat{\beta}_1\bar{x}_1-\left(f_0-1\right)\hat{\beta}_0\bar{x}_0+f_0\hat{\beta}_1\bar{x}_0-f_1\hat{\beta}_0\bar{x}_1\end{aligned}$$

But

$$f_1-1=\frac{N_1-\left(N_0+N_1\right)}{N_0+N_1}=-\frac{N_0}{N_0+N_1}=-f_0$$

and

$$f_0-1=\frac{N_0-\left(N_0+N_1\right)}{N_0+N_1}=-\frac{N_1}{N_0+N_1}=-f_1$$

Therefore,

$$\begin{aligned}\hat{\tau}_{ATE} &= \left(\bar{y}_1-\bar{y}_0\right)-f_0\hat{\beta}_1\bar{x}_1+f_1\hat{\beta}_0\bar{x}_0+f_0\hat{\beta}_1\bar{x}_0-f_1\hat{\beta}_0\bar{x}_1\\ &= \left(\bar{y}_1-\bar{y}_0\right)+\left(f_0\hat{\beta}_1+f_1\hat{\beta}_0\right)\bar{x}_0-\left(f_0\hat{\beta}_1+f_1\hat{\beta}_0\right)\bar{x}_1\\ &= \left(\bar{y}_1-\bar{y}_0\right)-\left(f_0\hat{\beta}_1+f_1\hat{\beta}_0\right)\left(\bar{x}_1-\bar{x}_0\right)\end{aligned}$$

CHAPTER 8

Heteroskedasticity

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the meaning of heteroskedasticity and give examples of data sets likely to exhibit heteroskedasticity.
2. Explain how and why plots of least squares residuals can reveal heteroskedasticity.
3. Specify a variance function and use it to test for heteroskedasticity with (a) a Breusch–Pagan test and (b) a White test.
4. Test for heteroskedasticity using a Goldfeld–Quandt test applied to (a) two subsamples with potentially different variances and (b) a model where the variance is hypothesized to depend on an explanatory variable.
5. Describe and compare the properties of the least squares and generalized least squares estimators when heteroskedasticity exists.
6. Compute heteroskedasticity-consistent standard errors for least squares.
7. Describe how to transform a model to eliminate heteroskedasticity.
8. Compute generalized least squares estimates for heteroskedastic models where (a) the variance is known except for the proportionality constant σ^2, (b) the variance is a function of explanatory variables and unknown parameters, and (c) the sample is partitioned into two groups with different variances.
9. Explain why the linear probability model exhibits heteroskedasticity.
10. Compute generalized least squares estimates of the linear probability model.

KEYWORDS

Breusch–Pagan test
generalized least squares
Goldfeld–Quandt test
grouped heteroskadasticity
heteroskedasticity
heteroskedasticity-consistent standard errors
homoskedasticity
Lagrange multiplier test
linear probability model
regression function
residual plot
robust standard errors
skedastic function
transformed model
variance function
weighted least squares
White test

8.1 The Nature of Heteroskedasticity

In Chapter 2, we discussed the relationship between household food expenditure and household income. We proposed the simple population regression model

$$FOOD\_EXP_i = \beta_1 + \beta_2 INCOME_i + e_i \tag{8.1}$$

Given the parameter values, β_1 and β_2, we can predict food expenditures for households with any income. Income is an important factor in households' decisions about weekly food expenditure, but there are many other factors entering a particular household's decisions. The random error e_i represents the collection of all the factors other than income that affect household expenditure on food.

The assumption of **strict exogeneity** says that when using information on household income our best prediction of the random error is zero. If sample values are randomly selected, then the technical expression for this assumption is that given income the conditional expected value of the random error e_i is zero, $E(e_i|INCOME_i) = 0$. If the assumption of strict exogeneity holds then the regression function is

$$E(FOOD\_EXP_i|INCOME_i) = \beta_1 + \beta_2 INCOME_i$$

The slope parameter β_2 describes how expected (population mean, or average) household food expenditure changes when household income increases by \$100, holding all else constant. The intercept parameter β_1 measures average expenditure on food for a household with no income in a week.

The discussion above focuses on the level, or amount, of food expenditure. We now ask, "How much **variation** in household food expenditure is there at different levels of income?" The U.S. median household income is about \$1000 a week. For such a household, the expected weekly food expenditure is $E(FOOD\_EXP_i|INCOME = 10) = \beta_1 + \beta_2(10)$. If we observe many households with the median income, we would observe a wide range of actual weekly food expenditures. The variation arises because different households have differing tastes and preferences, and they have differing demographic characteristics, and life circumstances. Readers who are students, and living on typical student incomes, how much variation is there in your food expenditure from week to week? We suspect that regardless of your tastes and preferences you have calculated very carefully how much you can afford and stick closely to a spending plan each week. In general, households with low incomes have little scope for wide variations in food expenditures from week to week because of their income constraint. On the other hand, households with a large weekly income have more food choices. Some high-income households may choose champagne, caviar, and steaks, but others may choose beer, rice, pasta, and beans. We can expect to observe larger variations in weekly food expenditures by households with large incomes.

Holding income constant, and given our model, what is the source of the variation in household food expenditures? It must be from the random error, the collection of factors, other than income, that influence food expenditure. As we observe different households at a given level of income, there are variations in food expenditures because randomly sampled households have different tastes and preferences and differ in many other ways as well. Recall that the random error in the regression is the difference between any observation on the outcome variable and its conditional expectation, that is

$$e_i = FOOD\_EXP_i - E(FOOD\_EXP_i|INCOME_i) \tag{8.2}$$

If the assumption of strict exogeneity holds, then the population average value of the random errors is $E(e_i|INCOME_i) = E(e_i) = 0$. A positive random error corresponds to an observation in which food expenditure is greater than expected, while a negative random error corresponds to an observation in which food expenditure is less than expected.

Another way of describing the greater variation in food expenditures for high-income households is to say the probability of observing large positive or negative random errors is higher

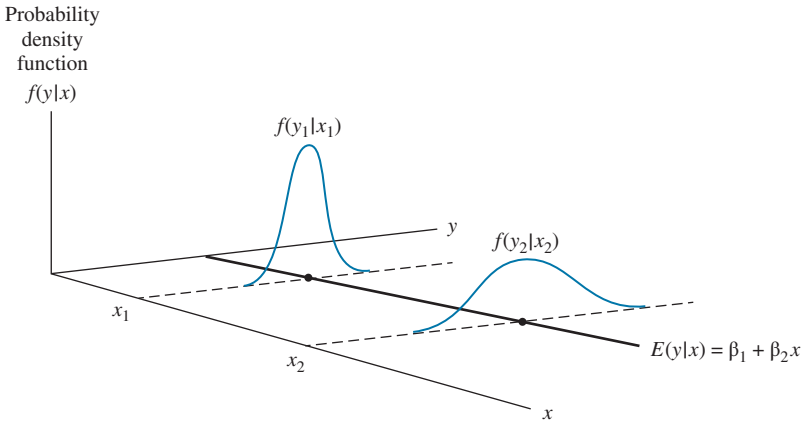


FIGURE 8.1 **Heteroskedastic errors.**

for high incomes than it is for low incomes. To illustrate this idea, examine Figure P.5 in the Probability Primer. First, suppose the probability distribution of the random errors is $N(0,1)$, the solid curve. What is the probability of observing a random value of e_i greater than two? Using Statistical Table 1, $P(e_i > 2) = P(Z > 2) = 0.0228$. Now, suppose the probability distribution of the random errors is $N(0, 4)$, the dot-dash curve. What is the probability of observing a random value of e_i greater than 2? Using Statistical Table 1, $P(e_i > 2) = P(Z > 1) = 0.1587$. The random error e_i has a higher probability of taking on a large value if its variance is large. In the context of the food expenditure example, we can capture the effect we are describing by assuming that $\text{var}(e_i|INCOME_i)$ increases as income increases. Food expenditure can deviate further from its mean, or expected value, when income is large.

In such a case, when the error variances for all observations are not the same, we say that **heteroskedasticity** exists. Alternatively, we say the random error e_i is **heteroskedastic**. Conversely, if all observations come from probability density functions with the same variance, we say that **homoskedasticity** exists, and e_i is **homoskedastic**. Heteroscedastic, homoscedastic, and heteroscedasticity are commonly used alternative spellings.

Figure 8.1 illustrates the heteroskedastic assumption. Let $y_i = FOOD\_EXP_i$ and $x_i = INCOME_i$. At x_1, the food expenditure probability density function $f(y_1|x_1)$ is such that y_1 will be close to $E(y_1|x_1)$ with high probability. When we move to the larger value x_2, the probability density function $f(y_2|x_2)$ is more spread out; we are less certain about where y_2 might fall, and much larger or smaller values than the average $E(y_2|x_2)$ are possible. When homoskedasticity exists, the probability density function for the errors does not change as x changes, as we illustrated in Figure 2.3.

8.2 Heteroskedasticity in the Multiple Regression Model

The existence of heteroskedasticity is a violation of one of our least squares assumptions listed in Section 5.1. For the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$, $i = 1, \ldots, N$, assumption MR3 is

$$\text{var}(e_i|\mathbf{X}) = \text{var}(y_i|\mathbf{X}) = \sigma^2$$

the conditional variance of the random error, and the dependent variable, is σ^2, a constant. Assumption MR3 is that the random error term is conditionally homoskedastic. The simplest

statement of the conditional heteroskedasticity assumption is

$$\text{var}(e_i|\mathbf{X}) = \text{var}(y_i|\mathbf{X}) = \sigma_i^2 \tag{8.3}$$

The change is very subtle, the error variance σ_i^2 now has a subscript, i, indicating that it is not always the same constant and may change from observation to observation, $i = 1, \ldots, N$. At the extreme, the error is heteroskedastic even if only one random error has a variance different than the other $N - 1$ random errors. Generally, however, we think of the problem as being more pervasive when it is present.

Assumptions MR1–MR5 apply to any type of regression, using time-series or cross-sectional data. Our notation $\mathbf{X}$ represents all N observations on $K - 1$ explanatory variables plus a constant term. Heteroskedasticity often arises when using **cross-sectional data**. The term cross-sectional data refers to having data on a number of economic units such as firms or households, *at a given point in time*. The household data on income and food expenditure fall into this category. Other possible examples include data on costs, outputs, and inputs for a number of firms, and data on quantities purchased and prices for some commodity, or commodities, in a number of retail establishments. Cross-sectional data usually involve observations on economic units of varying sizes. For example, data on households will involve households with varying numbers of household members and different levels of household income. With data on a number of firms, we might measure the size of the firm by the quantity of output it produces. Frequently, the larger the firm, or the larger the household, the more difficult it is to explain the variation in some outcome variable y_i by the variation in a set of explanatory variables. Larger firms and households are likely to be more diverse and flexible with respect to the way in which values for y_i are determined. What this means for the linear regression model is that, as the size of the economic unit becomes larger, there is more uncertainty associated with the outcomes y_i. We model this greater uncertainty by specifying a conditional error variance that is larger, the larger the size of the economic unit.

Heteroskedasticity is not a property that is necessarily restricted to cross-sectional data. With time-series data, where we have data over time on an economic unit, such as a firm, a household, or even a whole economy, it is possible that the conditional error variance will change. This would be true if there was an external shock or change in circumstances that created more or less uncertainty about y.

For simplification, in the remainder of this chapter, we assume that the errors are uncorrelated and that heteroskedasticity is an observation-by-observation problem and that the conditional variance of the ith observation's random error e_i is unrelated to the jth observation. In the context of the cross-sectional data food expenditure example, we are ruling out the case in which the variability in the random error component for the ith household is connected to or explained by the characteristics of the jth household. In a time-series regression context, we are ruling out the case when the error variation at time t is related to conditions in the past, at time $t - s$. Can we always rule out these exceptions? No, we cannot. In the cross-sectional data context, we may find that households drawn from some geographical regions, or neighborhoods, are similar, so that the error variation for neighboring households might be similar, or connected. In the time-series context, we most certainly cannot rule out continuous periods of stability, perhaps many weeks at a time, and periods of instability that can similarly last many weeks or months, meaning that the error variation at time t is related to the error variation at times $t - 1$, $t - 2$, and so on. For now, however, we will rule out these interesting cases.

8.2.1 The Heteroskedastic Regression Model

The multiple regression model is $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$. We assume we have a random sample so that the ith observation is statistically independent of the jth observation. Let $x_i = (1, x_{i2}, \ldots, x_{iK})$ denote the values of the K explanatory variables for the ith observation. The heteroskedasticity assumption in (8.3) becomes

$$\text{var}(y_i|\mathbf{x}_i) = \text{var}(e_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma_i^2 \tag{8.4}$$

where $h(\mathbf{x}_i) > 0$ is a function of $\mathbf{x}_i$ that is sometimes called the **skedastic function**,[1] and $\sigma^2 > 0$ is a constant. If $h(\mathbf{x}_i) = 1$, then the conditional variance is homoskedastic. If $h(\mathbf{x}_i)$ is not constant, then the conditional variance is heteroskedastic. For example, when $h(\mathbf{x}_i) = x_{ik}$ the conditional variance becomes $\text{var}(e_i|\mathbf{x}_i) = \sigma^2 x_{ik}$, the error variance is proportional to the kth explanatory variable x_{ik}. Because variances must be positive, for the proportional heteroskedasticity model to work $h(\mathbf{x}_i) = x_{ik} > 0$. In (8.4) we assume the conditional variance depends on the values of some or all of the explanatory variables in the regression equation.

This chapter is concerned with the consequences of a variance assumption like (8.4). What are the consequences for the properties of least squares estimator? Is there a better estimation technique? How do we detect the existence of heteroskedasticity?

EXAMPLE 8.1 | Heteroskedasticity in the Food Expenditure Model

We can further illustrate the nature of heteroskedasticity and at the same time demonstrate an informal way of detecting heteroskedasticity using the food expenditure data. Using the $N = 40$ observations in the data file *food,* the OLS estimates are

$$\widehat{FOOD\_EXP}_i = 83.42 + 10.21\,INCOME_i$$

A graph of this fitted line, along with all the observed expenditure–income points, appears in Chapter 2, Figure 2.8. Notice that, as income grows, the prevalence of data points that deviate further from the estimated mean function increases. There are more points scattered further away from the line as income gets larger. Another way of describing this feature is to say that there is a tendency for the least squares residuals, defined by

$$\hat{e}_i = FOOD\_EXP_i - 83.42 - 10.21 INCOME_i$$

to increase in absolute value as income grows. The plot of the absolute value of the residuals, $|\hat{e}_i|$, versus income in Figure 8.2 shows this quite clearly. The plot of the calculated residuals, $\hat{e}_i$, versus income in Figure 8.3 shows the characteristic "spray" pattern shown in Chapter 4, Figure 4.7(b). Figure 4.7(a) shows the random scatter we anticipate if the errors are conditionally homoskedastic. Figures 4.7(b)–(d), spray, funnel, and bowtie, are patterns we might observe when the errors are conditionally heteroskedastic.

Since the observable least squares residuals $(\hat{e}_i)$ are the analogues of the unobservable errors (e_i), Figures 8.2 and 8.3 also suggest that the unobservable errors tend to increase in absolute value as income increases. That is, the variation of food expenditure around the conditional mean food expenditure $E(FOOD\_EXP_i|INCOME_i) = \beta_1 + \beta_2 INCOME_i$, and variation in the random error term, increase as income increases. The conditional variance $\text{var}(e_i|INCOME_i) = \sigma^2 h(INCOME_i)$, where $h(INCOME_i)$ is an increasing function of income. Possible variance functions include

$$\text{var}(e_i|INCOME_i) = \sigma^2 INCOME_i$$

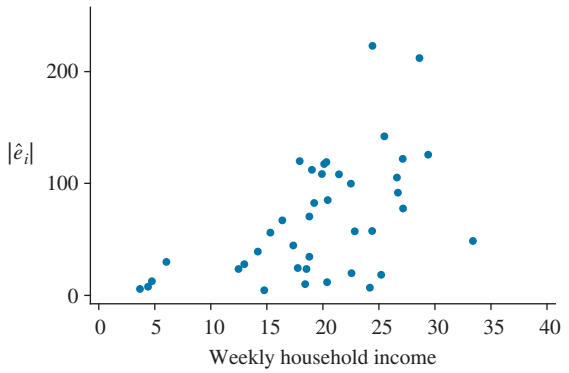


FIGURE 8.2 **Absolute value of food expenditure residuals vs. income.**

[1] See A. Colin Cameron and Pravin K. Trivedi (2010) *Microeconometrics Using Stata, Revised Edition*, Stata Press, p. 153.

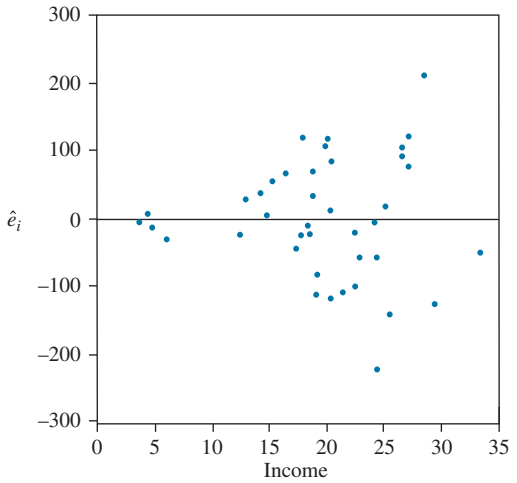


FIGURE 8.3 **Least squares food expenditure residuals plotted against income.**

or

$$\text{var}(e_i|INCOME_i) = \sigma^2 INCOME_i^2$$

These are consistent with the hypothesis that we posed earlier, namely, that the mean food expenditure function is better at explaining food expenditure for low-income households than it is for high-income households.

Plotting of least squares residuals is an informal way of detecting heteroskedasticity. Later in the chapter, in Section 8.6, we consider formal test procedures. First, however, we examine the consequences of heteroskedasticity for least squares estimation.

8.2.2 Heteroskedasticity Consequences for the OLS Estimator

Since the existence of heteroskedasticity violates the usual least squares assumption $\text{var}(e_i|\mathbf{x}_i) = \sigma^2$, we need to ask what consequences this violation has for our least squares estimator, and what we can do about it. There are two implications:

1. The least squares estimator is still a linear and unbiased estimator, but it is no longer best. There is another estimator with a smaller variance.
2. The standard errors usually computed for the least squares estimator are incorrect. Confidence intervals and hypothesis tests that use these standard errors may be misleading.

Let's first consider the simple linear regression model with homoskedasticity

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ with } \text{var}(e_i|\mathbf{x}) = \sigma^2 \tag{8.5}$$

We showed in Chapter 2 that the conditional variance of the least squares estimator for b_2 is

$$\text{var}(b_2|\mathbf{x}) = \sigma^2 \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{8.6}$$

Now suppose the error variances for each observation are different and that we recognize this difference by putting a subscript i on σ^2, so that we have

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ with } \text{var}(e_i|\mathbf{x}) = \sigma_i^2 \tag{8.7}$$

In Appendix 8A, we show that under the heteroskedastic specification in (8.7) the least squares estimator is unbiased with conditional variance

$$\text{var}(b_2|\mathbf{x}) = \left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{-1}\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\sigma_i^2\right]\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{-1} \tag{8.8}$$

If the errors are homoskedastic, then equation (8.8) reduces to the usual OLS estimator variance in equation (8.6). If the errors are heteroskedastic then (8.8) is correct and (8.6) is not. This is a practical problem because your computer software has programmed into it the estimated variances and covariance of the least squares estimator under homoskedasticity, given in Chapter 2, equations (2.20)–(2.22). This in turn means that if the errors are heteroskedastic, the usual standard errors in equations (2.23)–(2.24) are incorrect. Using incorrect standard errors in t-tests and confidence intervals may lead us to faulty conclusions. If we proceed to use the least squares estimator and its usual standard errors when $\text{var}(e_i) = \sigma_i^2$, we will be using an estimate of (8.6) to compute the standard error of b_2 when we should be using an estimate of (8.8).

8.3 Heteroskedasticity Robust Variance Estimator

Calculation of a correct estimate for the OLS variance (8.8) is astonishingly simple, although the theory leading to it is not. Simply replace σ_i^2 by $[N/(N-2)]\,\hat{e}_i^2$, the squared OLS residuals multiplied by an inflation factor.[2] The **White heteroskedasticity-consistent estimator (HCE)** that is valid in large samples for the simple regression model is

$$\widehat{\text{var}}(b_2) = \left[\sum(x_i-\bar{x})^2\right]^{-1}\left\{\sum\left[(x_i-\bar{x})^2\left(\frac{N}{N-2}\right)\hat{e}_i^2\right]\right\}\left[\sum(x_i-\bar{x})^2\right]^{-1} \tag{8.9}$$

where $\hat{e}_i$ is the least squares residual from the regression model, $y_i = \beta_1 + \beta_2 x_i + e_i$. The estimator is named after econometrician Halbert White who developed the idea. This variance estimator is **robust** because it is valid whether heteroskedasticity is present or not. Thus, if we are not sure whether the random errors are heteroskedastic or homoskedastic, then we can use a robust variance estimator and be confident that our standard errors, t-tests, and interval estimates are valid in large samples.

The formula in equation (8.9) has a lovely symmetry and is one illustration of a **variance sandwich**. Let $C = \left[\sum(x_i-\bar{x})^2\right]^{-1}$ be the "outside ***C***rust" and let $A = \left\{\sum\left[(x_i-\bar{x})^2\left(\frac{N}{N-2}\right)\hat{e}_i^2\right]\right\}$ be "***A***ny filling." Then our variance sandwich is *any filling* between two *crusts*, or $\widehat{\text{var}}(b_2) = CAC$. Modern Econometrics offers many such sandwiches. Equations (8.8) and (8.9) can be simplified, but we prefer to leave them as is to emphasize the "sandwich" form. Also the matrix approaches to multiple regression in your future econometrics courses will use the sandwich form.

EXAMPLE 8.2 | Robust Standard Errors in the Food Expenditure Model

Most regression packages include an option for calculating standard errors using White's estimator. If we do so for the food expenditure example, we obtain

$$\widehat{FOOD\_EXP} = 83.42 + 10.21\,INCOME$$
$$(27.46)\quad (1.81)\quad \text{(White robust se)}$$
$$(43.41)\quad (2.09)\quad \text{(incorrect OLS se)}$$

In this case, ignoring heteroskedasticity and using incorrect standard errors, based on the usual formula in (8.6), tends to understate the precision of estimation; we tend to get confidence intervals that are wider than they should be. Specifically, following the result in (3.6) in Chapter 3, we can construct corresponding 95% confidence intervals for β_2.

[2] See Appendix 8C for the logic of this inflation, and development of other versions of the robust variance.

White Robust se:

$$b_2 \pm t_c \text{se}(b_2) = 10.21 \pm 2.024 \times 1.81 = [6.55,\ 13.87]$$

Incorrect OLS se:

$$b_2 \pm t_c \text{se}(b_2) = 10.21 \pm 2.024 \times 2.09 = [5.97,\ 14.45]$$

If we ignore heteroskedasticity, we estimate that β_2 lies between 5.97 and 14.45. When we recognize the existence of heteroskedasticity, our information is judged more precise, and using the **robust standard error** we estimate that β_2 lies between 6.55 and 13.87. A caveat here is that the sample is small, which does mean that the robust standard error formula we have provided may not be as accurate as if the sample were large.

White's estimator for the standard errors helps us avoid computing incorrect interval estimates or incorrect values for test statistics in the presence of heteroskedasticity. However, it does not address the first implication of heteroskedasticity that we mentioned at the beginning of this section, that the least squares estimator is no longer best. However, failing to use the "best" estimator may not be too grave a sin if estimates are sufficiently precise for useful economic analysis. Many cross-sectional data sets have thousands of observations, resulting in robust standard errors that are small, making interval estimates narrow and *t*-tests powerful. Nothing further is required in these cases. If, however, your estimates are not sufficiently precise for economic analysis, then a better, more efficient, estimator is called for. In order to use such an estimator we must specify the **skedastic** function $h(\mathbf{x}_i) > 0$, a function of $\mathbf{x}_i$ and also perhaps other variables, that describes the pattern of conditional heteroskedasticity. In the next section, we describe an alternative estimator that has a smaller variance than the least squares estimator.

8.4 Generalized Least Squares: Known Form of Variance

To begin, consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. Let's assume the data are obtained by random sampling, so that the observations are statistically independent of one another, that $E(e_i|x_i) = 0$, and that the heteroskedasticity assumption is

$$\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma_i^2 \tag{8.10}$$

Although it is possible to obtain the White heteroskedasticity-consistent variance estimates by simply assuming the error variances σ_i^2 can be different for each observation, to develop an estimator that is better than the least squares estimator, we need to make a further assumption about how the variances σ_i^2 change with each observation. This means making an assumption about the skedastic function $h(x_i)$. The further assumption is necessary because the best linear unbiased estimator in the presence of heteroskedasticity, an estimator known as the **generalized least squares (GLS)** estimator, depends on the unknown σ_i^2. It is not practical to estimate N unknown variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$ with only N observations without making a restrictive assumption about how the σ_i^2 change. Thus, to make the GLS estimator operational some structure is imposed on σ_i^2. Alternative structures are considered in this and the following section. Details of the GLS estimator and the issues involved will become clear as we work our way through these sections.

8.4.1 Transforming the Model: Proportional Heteroskedasticity

Recall our earlier inspection of the least squares residuals for the food expenditure example. The variation in the OLS residuals increases as income increases, which suggests that the error

variance increases as income increases. One possible assumption for the variance σ_i^2 that has this characteristic is

$$\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i, \quad x_i > 0 \tag{8.11}$$

That is, we assume that the variance of the ith error term σ_i^2 is given by a positive unknown constant parameter σ^2 multiplied by the positive income variable x_i, so that $\text{var}(e_i|x_i)$ is **proportional** to income. We are assuming the skedastic function is $h(x_i) = x_i$. As explained earlier, in economic terms this assumption implies that, for low levels of income (x_i), food expenditure (y_i) will be clustered closer to the **regression function** $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Expenditure on food for low-income households will be largely explained by the level of income. At high levels of income, food expenditures can deviate more from the regression function. This means that there are likely to be many other factors, such as specific tastes and preferences, that reside in the error term, and that lead to a greater variation in food expenditure for high-income households.

The least squares estimator is **not** the best linear unbiased estimator when the errors are heteroskedastic. Is there a best linear unbiased estimator under these circumstances? Yes there is! The approach is to **transform the model** into one with homoskedastic errors. Leaving the basic structure of the model intact, we turn the heteroskedastic error model into a homoskedastic error model. After the transformation, applying OLS to the **transformed model** gives a best linear unbiased estimator. These steps define the new GLS estimator.

Given the model of proportional heteroskedasticity in equation (8.11), begin by dividing both sides of the original model in (8.7) by $\sqrt{x_i}$

$$\frac{y_i}{\sqrt{x_i}} = \beta_1\left(\frac{1}{\sqrt{x_i}}\right) + \beta_2\left(\frac{x_i}{\sqrt{x_i}}\right) + \frac{e_i}{\sqrt{x_i}} \tag{8.12}$$

Define the **transformed variables** and **transformed error** as

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_{i1}^* = \frac{1}{\sqrt{x_i}}, \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}} = \sqrt{x_i}, \quad e_i^* = \frac{e_i}{\sqrt{x_i}} \tag{8.13}$$

so that (8.12) can be rewritten as

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^* \tag{8.14}$$

The beauty of this transformed model is that the new transformed error term e_i^* is homoskedastic. To see this, recall equation (P.14) from the Probability Primer: If X is a random variable and a is a constant, then $\text{var}(aX) = a^2\text{var}(X)$. Applying that rule here we have

$$\text{var}(e_i^*|x_i) = \text{var}\left(\frac{e_i}{\sqrt{x_i}}\middle| x_i\right) = \frac{1}{x_i}\text{var}(e_i|x_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2 \tag{8.15}$$

Using the rules of expected values, the transformed error term will retain a zero conditional mean $E(e_i^*|x_i) = 0$. As a consequence, we can apply OLS to the transformed variables, y_i^*, x_{i1}^*, and x_{i2}^* to obtain the best linear unbiased estimator for β_1 and β_2. Note that the transformed variables y_i^*, x_{i1}^*, and x_{i2}^* are easy to create. An important difference between the original and transformed models is that the transformed model no longer contains a constant term. In the original model, $x_{i1} = 1$. In the transformed model, the variable $x_{i1}^* = 1/\sqrt{x_i}$ is no longer constant. You will have to be careful to exclude the constant if your software automatically inserts one, but you can still proceed. The transformed model is linear in the unknown parameters β_1 and β_2. These are the original parameters that we are interested in estimating. They are unaffected by the transformation. In short, the transformed model is a linear model to which we can apply OLS estimation. The transformed model satisfies the conditions of the Gauss–Markov theorem, and the OLS estimators defined in terms of the transformed variables are BLUE.

To summarize, to obtain the best linear unbiased estimator for a model with heteroskedasticity of the type specified in equation (8.11), $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i$:

1. Calculate the transformed variables given in (8.13).
2. Use OLS to estimate the transformed model given in (8.14), yielding estimates $\hat{\beta}_1$ and $\hat{\beta}_2$.

The estimates obtained in this way are the GLS estimates.

The GLS estimator is BLUE if the model assumption of proportional heteroskedasticity is correct. Of course, we never know if our assumed skedastic function is correct or not. It is likely that a thoughtfully chosen transformation will reduce the model heteroskedasticity. If, however, the chosen transformation does not completely eliminate the heteroskedasticity, the GLS estimator is linear and unbiased but not best, and the standard errors from the transformed model estimation are incorrect. What then? Easy. Use White robust standard errors with the transformed data model to obtain valid (in large samples) standard errors. Doing so we will have striven for a more efficient estimator, but been cautious to present valid standard errors, *t*-stats, and interval estimates. We illustrate this strategy in Example 8.3.

8.4.2 Weighted Least Squares: Proportional Heteroskedasticity

One way of viewing the GLS estimator is as a **weighted least squares (WLS)** estimator. Recall that the OLS estimates are those values of β_1 and β_2 that minimize the sum of squared errors

$$S(\beta_1, \beta_2|y_i, x_i) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

The sum of squares function using the transformed data model (8.14) is

$$\begin{aligned} S(\beta_1, \beta_2|y_i, x_i) &= \sum_{i=1}^{N} (y_i^* - \beta_1 x_{i1}^* - \beta_2 x_{i2}^*)^2 = \sum_{i=1}^{N} \left(\frac{y_i}{\sqrt{x_i}} - \beta_1 \frac{1}{\sqrt{x_i}} - \beta_2 \frac{x_{i2}}{\sqrt{x_i}} \right)^2 \\ &= \sum_{i=1}^{N} \left[\frac{1}{\sqrt{x_i}} (y_i - \beta_1 - \beta_2 x_{i2}) \right]^2 \\ &= \sum_{i=1}^{N} \frac{(y_i - \beta_1 - \beta_2 x_{i2})^2}{x_i} \end{aligned} \tag{8.16}$$

The squared errors are *weighted* by $1/x_i$. Recall that our variance assumption is $\text{var}(e_i|x_i) = \sigma^2 x_i$. When x_i is smaller we are assuming the variance of the error is smaller and the data fall closer to the regression function. These data are *more informative* about the location of $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. When x_i is larger we are assuming the variance of the error is larger, and the data may fall farther from the regression function. These data are *less informative* about the location of $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Intuitively, it makes sense to "down weight" observations with less information and weigh more heavily observations with more information. That is exactly what the weighted sum of squares function (8.16) achieves. When x_i is small, the data contain more information about the regression function and the observations are weighted heavily. When x_i is large, the data contain less information and the observations are weighted lightly. In this way, we take advantage of the heteroskedasticity to improve parameter estimation. On the other hand, OLS estimation treats all observations as equally informative and equally important, as it should under homoskedasticity.

Most software have a WLS or GLS option. If your software falls into this category, you do not have to transform the variables before estimation, nor do you have to worry about omitting the constant. The computer will do both the transforming and the estimating once you decipher the software command. If you do the transforming yourself, that is, you create y_i^*, x_{i1}^*, and x_{i2}^*,

and apply OLS, be careful not to include a constant in the regression. As noted before, there is no constant because $x_{i1}^* \neq 1$.

EXAMPLE 8.3 | Applying GLS/WLS to the Food Expenditure Data

In the food expenditure example, we assume $\text{var}(e_i|INCOME_i) = \sigma_i^2 = \sigma^2 INCOME_i$. Applying the generalized (weighted) least squares procedure to our household expenditure data yields the following GLS estimates:

$$\underset{\text{(se)}}{\widehat{FOOD\_EXP}_i} = \underset{(23.79)}{78.68} + \underset{(1.39)}{10.45} INCOME_i \tag{8.17}$$

That is, we estimate the intercept term as $\hat{\beta}_1 = 78.68$ and the slope coefficient that shows the response of food expenditure to a change in income as $\hat{\beta}_2 = 10.45$. These estimates are somewhat different from the least squares estimates $b_1 = 83.42$ and $b_2 = 10.21$ that did not allow for the existence of heteroskedasticity. It is important to recognize that the interpretations for β_1 and β_2 are the same in the transformed model in (8.14) as they are in the untransformed model in (8.7). Transformation of the variables is a technique for converting a heteroskedastic error model into a homoskedastic error model, *not* as something that changes the meaning of the coefficients.

The standard errors in (8.17), $\text{se}(\hat{\beta}_1) = 23.79$ and $\text{se}(\hat{\beta}_2) = 1.39$, are both lower than their least squares counterparts that were calculated from White's robust standard errors, namely, $\text{se}(b_1) = 27.46$ and $\text{se}(b_2) = 1.81$. Since GLS is a better estimation procedure than least squares, we expect the GLS standard errors to be lower. This statement needs to be qualified in two ways. First, remember that standard errors are square roots of *estimated* variances; in a single sample, the relative magnitudes of true variances may not always be reflected by their corresponding variance estimates. Second, the reduction in variance has come at the cost of making an additional assumption, namely, that the error variances have the structure given in (8.11).

The smaller standard errors have the advantage of producing narrower, more informative confidence intervals. For example, using the GLS results, a 95% confidence interval for β_2 is given by

$$\hat{\beta}_2 \pm t_c \text{se}(\hat{\beta}_2) = 10.451 \pm 2.024 \times 1.386 = [7.65, 13.26]$$

The least squares confidence interval computed using White's standard errors was [6.55, 13.87].

In order to obtain the GLS estimates, we assumed the specific pattern of heteroskedasticity, namely $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i$. We must ask ourselves whether this assumption adequately represents the pattern of heteroskedasticity in the data. If so, then the transformed model (8.14) should have homoskedastic errors. An informal check is to compute the residuals from the transformed model and plot them. That is, let $\hat{e}_i^* = y_i^* - \hat{\beta}_1 x_{i1}^* - \hat{\beta}_2 x_{i2}^*$. If you have used a WLS/GLS software, then the residuals it saves are, most likely, the GLS residuals $\hat{e}_{i,\text{WLS}} = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2}$. In this case, $\hat{e}_i^* = \hat{e}_{i,\text{WLS}}/\sqrt{x_i}$. In Figure 8.4 we plot the residuals from the transformed model and the OLS residuals against household income.

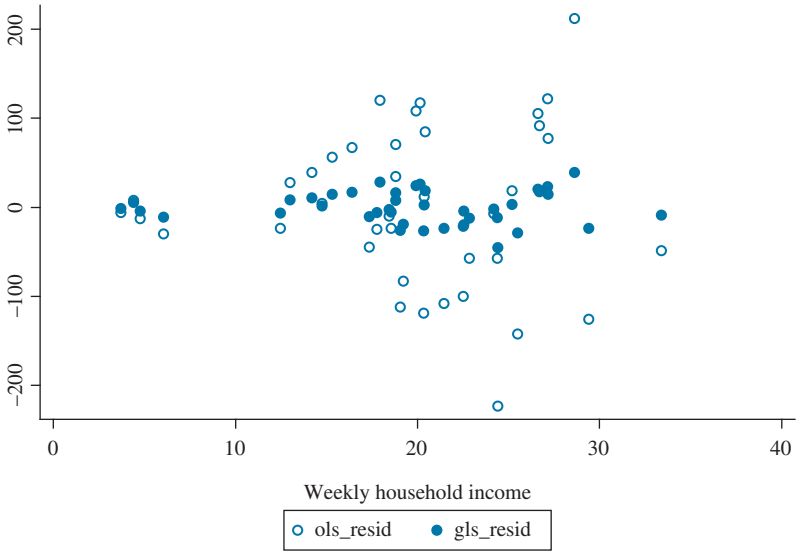


FIGURE 8.4 **OLS- and GLS-transformed residuals.**

It is evident that our transformation has substantially reduced the "spray" pattern indicating heteroskedasticity. If the transformation is a total success plotting the transformed residuals against *any* variable should reveal no pattern. If patterns remain, then you may try another skedastic function. Or, because it is visually clear that the transformation eliminated most, if not all, the heteroskedasticity, we can use a White heteroskedasticity robust standard error with the transformed model. In this way, we will have attempted to gain a more efficient estimator, but then protected ourselves against incorrect standard errors from any remaining heteroskedasticity. The GLS/WLS estimated model with robust standard errors is

$$\underset{\text{(robse)}}{\widehat{FOOD\_EXP}_i} = \underset{(12.04)}{78.68} + \underset{(1.17)}{10.45}INCOME_i$$

The 95% interval estimate of the slope is [8.07, 12.83].

8.5 Generalized Least Squares: Unknown Form of Variance

In the previous section, we assumed that heteroskedasticity could be described by the **variance function** $\text{var}(e_i|x_i) = \sigma^2 x_i$. This is convenient and simple in the food expenditure example because $x_i = INCOME_i > 0$ and intuitively reasonable. However, this is one possible choice of a skedasticity function $h(x_i)$. There are other alternatives such as $\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma^2 x_i^2$ and $\text{var}(e_i|x_i > 0) = \sigma^2 h(x_i) = \sigma^2 x_i^{1/2}$. Both have the property that the error variance increases as x_i increases. Why not choose one of these functions?

In a multiple regression $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ a heteroskedasticity pattern might be related to more than one of the explanatory variables, so that we might consider a skedastic function $h(x_{i2}, \ldots, x_{iK}) = h(\mathbf{x}_i)$. In fact, the heteroskedasticity pattern might be related to variables not even in the model! In order to deal with the more general specification that includes all these possibilities we need a model that is flexible, parsimonious, and for which $\sigma_i^2 > 0$. One specification that works well is

$$\begin{aligned}\sigma_i^2 &= \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \exp(\alpha_1)\exp(\alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \sigma^2 h(z_{i2}, \ldots, z_{iS})\end{aligned} \tag{8.18}$$

The candidate variables $z_{i2}, \ldots, z_{iS}$ that are possibly associated with the heteroskedasticity may or may not be in $\mathbf{x}_i$. The exponential function is convenient because it ensures we will get positive values for the variances σ_i^2 for all possible values of the parameters $\alpha_1, \alpha_2, \ldots, \alpha_S$. Equation (8.18) is called the model of **multiplicative heteroskedasticity**. It includes homoskedasticity as a special case; when $\alpha_2 = \cdots = \alpha_S = 0$ the error variance is $\sigma_i^2 = \exp(\alpha_1) = \sigma^2$. It is called a multiplicative model because

$$\exp(\alpha_1)\exp(\alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) = \exp(\alpha_1)\exp(\alpha_2 z_{i2}) \cdots \exp(\alpha_S z_{iS})$$

Each candidate variable has a separate multiplicative effect. This model does introduce some new parameters, but as you have seen many times now, when there is an unknown parameter an econometrician will figure out how to estimate it. That is what we do.

This model is attractive because of the features mentioned above, it is flexible, parsimonious, and $\sigma_i^2 > 0$, and also because it has several special cases that are very useful.

Multiplicative Heteroskedasticity, Special Case 1: $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha_2}$

As noted in the food expenditure example, three plausible variance functions are $\text{var}(e_i|x_i) = \sigma^2 x_i$, $\text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma^2 x_i^2$, and $\text{var}(e_i|x_i > 0) = \sigma^2 h(x_i) = \sigma^2 x_i^{1/2}$. These are special cases of

$$\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha_2}$$

where α_2 is an unknown parameter. In the multiplicative model, let $S = 2$, $z_{i2} = \ln(x_i)$ and $h(z_{i2}) = \exp[\alpha_2 \ln(x_i)]$. Using the properties of logarithms and exponentials, we have

$$\begin{aligned}\sigma_i^2 &= \exp(\alpha_1 + \alpha_2 z_{i2}) \\ &= \exp(\alpha_1)\exp\left[\alpha_2 \ln(x_i)\right] = \exp(\alpha_1)\exp\left[\ln(x_i^{\alpha_2})\right] \\ &= \sigma^2 x_i^{\alpha_2}\end{aligned}$$

Multiplicative Heteroskedasticity, Special Case 2: Grouped Heteroskedasticity

Data partitions arise naturally in many economic examples. We might be estimating a wage equation with data on individuals from both urban and rural areas. It is likely that the labor market in the urban area is more diverse, leading to wage variations from one person to another that is greater than in a rural area. Or perhaps we are considering wages for individuals with different education levels, such as those with only primary school education, those with a high school education, and those with some postsecondary education. Or individuals in different industries, or countries, etc. It is possible that the same basic structure holds for each group, with perhaps intercept dummy variables, and an error variance that is different for one group versus another.

Suppose we are considering just two groups. Create an indicator variable $D_i = 1$ if an observation is in one group and $D_i = 0$ for observations in the other group. Then the variance function is

$$\text{var}(e_i|\mathbf{x}_i) = \exp(\alpha_1 + \alpha_2 D_i) = \begin{cases} \exp(\alpha_1) = \sigma^2 & D_i = 0 \\ \exp(\alpha_1 + \alpha_2) = \sigma^2 \exp(\alpha_2) & D_i = 1 \end{cases}$$

Using the multiplicative form $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 D_i) = \exp(\alpha_1)\exp(\alpha_2 D_i) = \sigma^2 h(D_i)$, the skedastic function is $h(D_i) = \exp(\alpha_2 D_i)$. Note that if $\alpha_2 = 0$ the error variance is the same for the two groups, meaning that the assumption of homoskedasticity holds.

The same strategy works if there are more than two groups. Suppose there are $g = 1, 2, \ldots, G$ groups or data partitions. Create indicator variables for each group. Let $D_{ig} = 1$ if an observation is from group g, and otherwise $D_{ig} = 0$. If e_{ig} is the random error for the ith observation in group g, then a useful variance function is

$$\text{var}(e_{ig}|\mathbf{x}_{ig}) = \exp(\alpha_1 + \alpha_2 D_{i2} + \cdots + \alpha_G D_{iG}) = \begin{cases} \exp(\alpha_1) = \sigma^2 = \sigma_1^2 & g = 1; \text{ only } D_{i1} = 1 \\ \exp(\alpha_1 + \alpha_2) = \sigma_2^2 & g = 2; \text{ only } D_{i2} = 1 \\ \vdots & \\ \exp(\alpha_1 + \alpha_G) = \sigma_G^2 & g = G; \text{ only } D_{iG} = 1 \end{cases}$$

In this specification, we have chosen group 1 as the reference group and its indicator variable is omitted. This is similar to the indicator variable approach in Chapter 7. The variance of the reference group error can be denoted σ^2 or σ_1^2, to indicate that it is for group 1. For groups 2, … , G the skedastic function is $h(D_g) = \exp(\alpha_g D_g)$. Alternatively, let the variance function be $\text{var}(e_{ig}|\mathbf{x}_{ig}) = \exp(\alpha_1 D_{i1} + \alpha_2 D_{i2} + \cdots + \alpha_G D_{iG})$. Work out the variance for each group with this alteration. The end results using these two specifications are identical.

8.5.1 Estimating the Multiplicative Model

How do we proceed with estimation with an assumption like (8.18)? Our ultimate objective is to estimate the regression parameters $\beta_1, \beta_2, \ldots, \beta_K$. With the model of multiplicative heteroskedasticity, we use several estimation steps.

FEASIBLE GLS PROCEDURE

1. Estimate the original model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ by OLS, saving the OLS residuals $\hat{e}_i$.
2. Use the least squares residuals and the variables $z_{i2}, \ldots, z_{iS}$ to estimate $\alpha_1, \alpha_2, \ldots, \alpha_S$.
3. Calculate the estimated skedastic function $\hat{h}(z_{i2}, \ldots, z_{iS})$.
4. Divide each observation by $\sqrt{\hat{h}(z_{i2}, \ldots, z_{iS})}$ and apply OLS to the transformed data, or use WLS regression with weighting factor $1/\hat{h}(z_{i2}, \ldots, z_{iS})$.

The resulting estimates, $\hat{\hat{\beta}}_1, \hat{\hat{\beta}}_2, \ldots, \hat{\hat{\beta}}_K$, are called **feasible generalized least squares (FGLS)** estimates or **estimated generalized least squares (EGLS)** estimates. If heteroskedasticity is present, the FGLS estimator is consistent and more efficient than OLS in large samples. We have placed a second "hat" on these estimates to differentiate them from the earlier GLS estimates and to remind us that these estimates depend on a first-stage estimation.

Step 2 in the procedure is accomplished through a very clever manipulation of the model of multiplicative heteroskedasticity. Taking logarithms of both sides of (8.18), we obtain

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}$$

This looks like a regression model except for the fact that the left-hand side is unknown. Add the log of the squared least squares residuals to each side:

$$\ln(\sigma_i^2) + \ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2) \tag{8.19}$$

Rearrange and simplify equation (8.19):

$$\begin{aligned}
\ln(\hat{e}_i^2) &= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2) - \ln(\sigma_i^2) \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln(\hat{e}_i^2/\sigma_i^2) \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + \ln\left[(\hat{e}_i/\sigma_i)^2\right] \\
&= \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i
\end{aligned}$$

We have taken the model of multiplicative heteroskedasticity and through some simple manipulations arrived to

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.20}$$

Using this model we can estimate $\alpha_1, \alpha_2, \ldots, \alpha_S$ in (8.19) using OLS and continue with the steps of the procedure. Whether or not this procedure is a legitimate one depends on the properties of the new error term v_i that we introduced in (8.20). Does it have a zero mean? Is it homoskedastic? In small samples the answer to these questions is no. However, in large samples the answer is happier. It can be shown (see Appendix 8C.1) that $E(v_i|\mathbf{z}_i) \cong -1.2704$ and $\text{var}(v_i|\mathbf{z}_i) \cong 4.9348$, where $\mathbf{z}_i = (1, z_{i2}, \ldots, z_{iS})$, and if $e_i \sim N(0, \sigma_i^2)$. Because the regression error does not have conditional

mean zero, the estimated value of α_1 will be off by -1.2704. But $\hat{\alpha}_2, \ldots, \hat{\alpha}_S$ are consistent estimators, which for estimating the skedastic function $\hat{h}(z_{i2}, \ldots, z_{iS})$ is all that matters.

EXAMPLE 8.4 | Multiplicative Heteroskedasticity in the Food Expenditure Model

In the food expenditure example, with z_{i2} defined as $z_{i2} = \ln(INCOME_i)$, the least squares estimate of (8.19) is

$$\widehat{\ln(e_i^2)} = 0.9378 + 2.329 \ln(INCOME_i)$$

Notice that the estimate $\hat{\alpha}_2 = 2.329$ is more than twice the value of $\alpha_2 = 1$, which was an implicit assumption of the variance specification used in Example 8.3. This suggests the earlier transformation was not sufficiently aggressive. Following the steps to obtain FGLS estimates we transform the model by dividing both sides by $\sqrt{\hat{h}(z_{i2})}$, where $\hat{h}(z_{i2}) = \exp[\hat{\alpha}_2 \ln(INCOME_i)]$, then apply OLS to the transformed data, or use WLS with weight $1/\hat{h}(z_{i2})$. The resulting FGLS estimates for the food expenditure example are

$$\underset{(\text{se})}{\widehat{FOOD\_EXP_i}} = \underset{(9.71)}{76.05} + \underset{(0.97)}{10.63} INCOME_I \qquad (8.21)$$

Compared to the GLS results for the variance specification $\sigma_i^2 = \sigma^2 INCOME_i$, the estimates for β_1 and β_2 have not changed a great deal, but there has been a considerable drop in the standard errors that, under the previous specification, were $\text{se}(\hat{\beta}_1) = 23.79$ and $\text{se}(\hat{\beta}_2) = 1.39$.

We must ask ourselves whether our FGLS transformation has been adequate; does the transformed model satisfy the homoskedasticity assumption? In Example 8.3, we computed the residuals from the transformed model $\hat{e}_i^* = y_i^* - \hat{\beta}_1 x_{i1}^* - \hat{\beta}_2 x_{i2}^*$. Similarly, let $\hat{\hat{e}}_i^* = y_i^{**} - \hat{\beta}_1 x_{i1}^{**} - \hat{\beta}_2 x_{i2}^{**}$, where $y_i^{**} = y_i / \sqrt{\hat{h}(z_{i2})}$, $x_{i1}^{**} = 1/\sqrt{\hat{h}(z_{i2})}$, and $x_{i2}^{**} = x_{i2}/\sqrt{\hat{h}(z_{i2})}$. In Figure 8.5, we plot $\hat{e}_i^*$ (empty circles) from the GLS-transformed model, and $\hat{\hat{e}}_i^*$ (solid dots), from the FGLS-transformed model, versus income. Note that the vertical axis scales in Figures 8.4 and 8.5 are different; so take that into account when comparing them. By "zooming in" on $\hat{e}_i^*$ (empty circles) from the GLS-transformed model, we see a fan-shaped pattern persisting, meaning that the GLS transformation did not completely eliminate heteroskedasticity. In Figure 8.4, we saw a great reduction in the "spray" pattern and in Figure 8.5 the FGLS-transformed model has yet smaller residuals and shows a further reduction in the "spray" pattern. Based on visual evidence, the FGLS model has done a better job at eliminating heteroskedasticity than the GLS model.

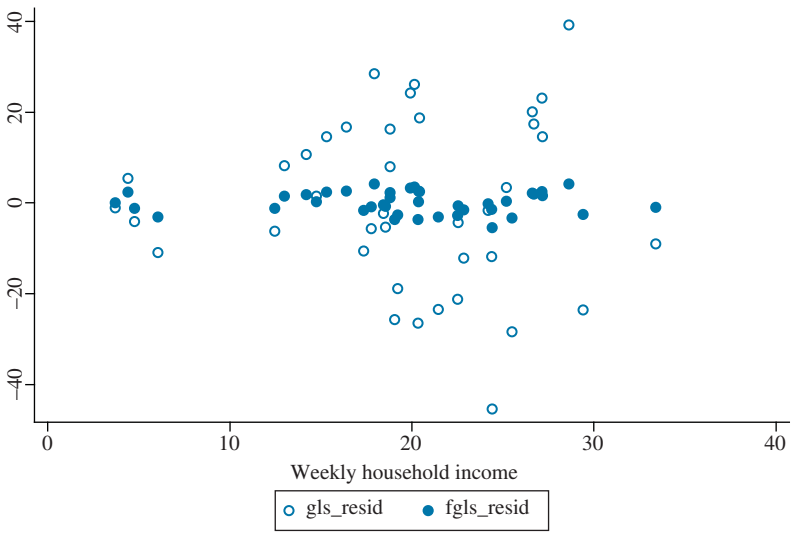


FIGURE 8.5 **GLS- and FGLS-transformed residuals.**

EXAMPLE 8.5 | A Heteroskedastic Partition

To illustrate the idea of a heteroskedastic partition we consider a simple wage equation in which a person's wage rate (*WAGE*) depends on their education (*EDUC*) and experience (*EXPER*). We also include an indicator variable for whether they live in a metropolitan, more urbanized, area or not. For convenience, think of the nonmetropolitan areas as "rural." That is

$$METRO = \begin{cases} 1 & \text{if person lives in a metropolitan area} \\ 0 & \text{if person lives in a rural area} \end{cases}$$

The wage equation is

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i$$

The issue we address here is the possibility that the variance of the error term is different in metropolitan areas than in rural areas. That is, we suspect that

$$\text{var}(e_i|\mathbf{x}_i) = \begin{cases} \sigma_M^2 & \text{if } METRO = 1 \\ \sigma_R^2 & \text{if } METRO = 0 \end{cases}$$

For illustration, we use the data file *cps5_small* and restrict ourselves to observations from the Midwest region, $MIDWEST = 1$. First consider the summary statistics in Table 8.1 for metropolitan workers, $METRO = 1$, and rural workers, $METRO = 0$.

Observe that the average wage and the standard deviation of wage are higher in metropolitan areas than in rural areas. This is suggestive but not proof of heteroskedasticity. The standard deviation is an "unconditional" measure that does not depend on the regression model. Heteroskedasticity is a concern about the variation in the regression random errors holding other factors constant, in this case education and experience.

The OLS estimates with heteroskedasticity robust standard errors are

$$\underset{\text{(robse)}}{\widehat{WAGE}_i} = \underset{(4.023)}{-18.450} + \underset{(0.261)}{2.339}EDUC_i + \underset{(0.0478)}{0.189}EXPER_i + \underset{(1.159)}{4.991}METRO_i$$

We save the OLS residuals, $\hat{e}_i$, and estimate equation (8.20) using $z_{i2} = METRO_i$, $\ln\left(\hat{e}_i^2\right) = \alpha_1 + \alpha_2 METRO + v_i$, obtaining

$$\widehat{\ln\left(\hat{e}_i^2\right)} = 2.895 + 0.700METRO$$

The estimated skedastic function is

$$\hat{h}(z_{i2}) = \exp(\hat{\alpha}_2 METRO_i) = \exp(0.700METRO) = \begin{cases} 2.0147 & METRO = 1 \\ 1 & METRO = 0 \end{cases}$$

We estimate the conditional variance of the random error to be about twice as large for the metropolitan area as in the rural area. In the WLS regression, the observations in the metropolitan area will receive half the weight of the observations in the rural area. The feasible GLS estimates are

$$\underset{\text{(se)}}{\widehat{\widehat{WAGE}}_i} = \underset{(3.788)}{-16.968} + \underset{(0.239)}{2.258}EDUC_i + \underset{(0.0447)}{0.175}EXPER_i + \underset{(1.214)}{4.995}METRO_i$$

The FGLS coefficient estimates and standard errors for *EDUC* and *EXPER* are slightly smaller than in the OLS estimation.

TABLE 8.1 Summary Statistics, by *METRO*

	Variable	Obs	Mean	Std. Dev.
$METRO = 1$	*WAGE*	213	24.25	14.00
	EDUC	213	14.25	2.77
	EXPER	213	23.15	13.17
$METRO = 0$	*WAGE*	84	18.86	8.52
	EDUC	84	13.99	2.26
	EXPER	84	24.30	14.32

8.6 Detecting Heteroskedasticity

In our discussion of the food expenditure equation, we used the nature of the economic problem and data to argue why heteroskedasticity of a particular form might be present. However, in many applications, there is uncertainty about the presence, or absence, of heteroskedasticity. It is natural to ask: How do I know if heteroskedasticity is likely to be a problem for my model and my set of data? Is there a way of detecting heteroskedasticity so that I know whether to use GLS techniques? We consider three ways of investigating these questions. The first is the informal use of **residual plots**. The other two are more formal classes of statistical tests.

8.6.1 Residual Plots

One way of investigating the existence of heteroskedasticity is to estimate your model using least squares and to plot the least squares residuals. If the errors are homoskedastic, there should be no patterns of any sort in the residuals, as shown in Figure 4.7(a). If the errors are heteroskedastic, they may tend to exhibit greater, or less, variation in some systematic way, as in Figures 4.7(b)–(d). For example, for the household food expenditure data, we suspect that the variance increases as incomes increases. We illustrated the use of diagnostic residual plots in Examples 8.1–8.3. We discovered that the absolute values of the residuals do indeed tend to increase as income increases. This method of investigating heteroskedasticity can be followed for any simple regression.

When we have more than one explanatory variable, the estimated least squares function is not so easily depicted on a diagram. However, what we can do is plot the least squares residuals against each explanatory variable, or against the fitted values $\hat{y}_i$, to see if those residuals vary in a systematic way relative to the specified variable.

8.6.2 The Goldfeld–Quandt Test

The second test for heteroskedasticity that we consider is designed for the case where we have two subsamples with possibly different variances. The sub-samples might be based on an indicator variable. In Example 8.5, we considered metropolitan and rural sub-samples for estimating a wage equation. Alternatively, we might sort the data according to the magnitude of one continuous variable and then divide the data into subsamples, omitting a few central observations to create separation if possible. In either case, the **Goldfeld–Quandt** test uses the estimated error variances from separate sub-sample regressions as a basis for the test. The background for this test appears in Appendix C.7.3. The only difference is in the degrees of freedom. Let the first sub-sample contain N_1 observations and let the regression model in this partition have K_1 parameters, including the intercept. Let the true variance of the error in this sample be σ_1^2 with estimator $\hat{\sigma}_1^2 = SSE_1/(N_1 - K_1)$. Let the second sub-sample contain N_2 observations and let the regression model in this partition have K_2 parameters, including the intercept. Let the true variance of the error in this sample be σ_2^2 with estimator $\hat{\sigma}_2^2 = SSE_2/(N_2 - K_2)$. The test statistic is

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(N_1-K_1,\, N_2-K_2)} \tag{8.22}$$

If the null hypothesis $H_0 : \sigma_1^2/\sigma_2^2 = 1$ is true, then the test statistic $GQ = \hat{\sigma}_1^2/\hat{\sigma}_2^2$ has an F-distribution with $(N_1 - K_1)$ numerator and $(N_2 - K_2)$ denominator degrees of freedom. If the alternative hypothesis is $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$, then we carry out a two-tail test. If we choose level of significance $\alpha = 0.05$, then we reject the null hypothesis if $GQ \geq F_{(0.975,\, N_1-K_1,\, N_2-K_2)}$ or if $GQ \leq F_{(0.025,\, N_1-K_1,\, N_2-K_2)}$, where $F_{(\alpha,\, N_1-K_1,\, N_2-K_2)}$ denotes the 100α-percentile of the F-distribution with the specified degrees of freedom. If the alternative is one-sided, $H_1 : \sigma_1^2/\sigma_2^2 > 1$, then we reject the null hypothesis if $GQ \geq F_{(0.95,\, N_1-K_1,\, N_2-K_2)}$.

EXAMPLE 8.6 | The Goldfeld–Quandt Test with Partitioned Data

We illustrate the Goldfeld–Quandt test by continuing Example 8.5. The data partitions are based on the indicator variable

$$METRO = \begin{cases} 1 & \text{if person lives in a metropolitan area} \\ 0 & \text{if person lives in a rural area} \end{cases}$$

The issue we address here is the possibility that the variance of the error term is different in metropolitan areas than in rural areas. To test the homoskedasticity assumption, estimate the wage equation in each data partition:

$$WAGE_{Mi} = \beta_{M1} + \beta_{M2} EDUC_{Mi} + \beta_{M3} EXPER_{Mi} + e_{Mi}$$
$$WAGE_{Ri} = \beta_{R1} + \beta_{R2} EDUC_{Ri} + \beta_{R3} EXPER_{Ri} + e_{Ri}$$

Let $\text{var}(e_{Mi}|\mathbf{x}_{Mi}) = \sigma_M^2$ and $\text{var}(e_{Ri}|\mathbf{x}_{Ri}) = \sigma_R^2$. Our null hypothesis is $H_0 : \sigma_M^2/\sigma_R^2 = 1$. Let the alternative hypothesis

be $H_1: \sigma_M^2/\sigma_R^2 \neq 1$, so that we use a two-tail test. The metropolitan subsample has 213 observations and the rural subsample has 84. In this case, as in most, the number of parameters in each data-partition regression is the same, $K = K_1 = K_2 = 3$. The test critical values are $F_{(0.975,\,210,\,81)} = 1.4615$ and $F_{(0.025,\,210,\,81)} = 0.7049$. Using $\widehat{\text{var}}(e_{Mi}|\mathbf{x}_{Mi}) = \hat{\sigma}_M^2 = 147.62$ and $\widehat{\text{var}}(e_{Ri}|\mathbf{x}_{Ri}) = \hat{\sigma}_R^2 = 56.71$, the calculated value of the Goldfeld–Quandt test statistic is $GQ = 2.6033 > F_{(0.975,\,210,\,81)} = 1.4615$, so we reject the null hypothesis that the error variances in the two subsamples are equal.

EXAMPLE 8.7 | The Goldfeld–Quandt Test in the Food Expenditure Model

Although the Goldfeld–Quandt test is very convenient for instances where the sample divides naturally into two subsamples, it can also be used where, under H_1, the variance is a function of a single explanatory variable. In the food expenditure model, we suspect that the error variance increases as income increases. We order the observations according to the magnitude of income so that, if heteroskedasticity exists, the first half of the sample will correspond to observations with lower variances and the last half of the sample will correspond to observations with higher variances. Then, we split the sample into two approximately equal halves, carry out two separate least squares regressions that yield variance estimates, say $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, and proceed with the test as described previously.

Following these steps for the food expenditure example, with the observations ordered according to income, we split the sample into two equal subsamples of 20 observations each. Because the sample is small, we do not omit any middle observations. Estimating the model on each subsample yields $\hat{\sigma}_1^2 = 3574.8$ and $\hat{\sigma}_2^2 = 12,921.9$, from which we obtain

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{12,921.9}{3574.8} = 3.61$$

Believing that the variances could increase, but not decrease with income, we use a one-tailed test with 5% level of significance critical value $F_{(0.95,\,18,\,18)} = 2.22$. Since $3.61 > 2.22$, a null hypothesis of homoskedasticity is rejected in favor of the alternative that the variance increases with income.

8.6.3 A General Test for Conditional Heteroskedasticity

In this section we consider a test for **conditional heteroskedasticity** that is related to some "explanatory" variables. Our equation of interest is the regression model

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{8.23}$$

Under assumptions MR1–MR5 the OLS estimator is the best linear unbiased estimator of the parameters $\beta_1, \beta_2, \ldots, \beta_K$. When conditional heteroskedasticity is a possibility, we hypothesize that the variance of the random error, e_i, depends on a set of explanatory variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ that may include some or all of the explanatory variables $x_{i2}, \ldots, x_{iK}$. That is, assume a general expression for the conditional variance

$$\text{var}(e_i|\mathbf{z}_i) = \sigma_i^2 = E(e_i^2|\mathbf{z}_i) = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \tag{8.24}$$

where $h(\cdot)$ is some smooth function and $\alpha_2, \alpha_3, \ldots, \alpha_S$ are **nuisance parameters**, meaning that we are not really interested in their values but must recognize that they are there. The beauty of the test we are about to present is that we do not have to actually know, or even guess, the function $h(\cdot)$. We will test for *any* relationship between the variance of the error term and *any* function of the selected variables. The function $h(\cdot)$ is similar to the skedastic function in equation (8.4), but here we have not factored out a constant σ^2, and unlike the feasible GLS estimation we do not have to choose an exponential form for $h(\cdot)$.

Notice what happens to the function $h(\cdot)$ when $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$. It collapses to

$$h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) = h(\alpha_1) \tag{8.25}$$

The term $h(\alpha_1)$, which we can define to be σ^2, is a constant, and $\text{var}(e_i|\mathbf{z}_i) = h(\alpha_1) = \sigma^2$. In other words, when $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$ the random errors are homoskedastic. On the other hand, if *any* of the parameters $\alpha_2, \alpha_3, \ldots, \alpha_S$ are not zero, then heteroskedasticity is present. Consequently, the null and alternative hypotheses for a test for heteroskedasticity based on the variance function are

$$\begin{aligned} &\text{homoskedasticity} \leftrightarrow H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0 \\ &\text{heteroskedasticity} \leftrightarrow H_1 : \text{ not all the } \alpha_s \text{ in } H_0 \text{ are zero} \end{aligned} \tag{8.26}$$

The null and alternative hypotheses are the first components of a test. The next component is a test statistic. To obtain a test statistic, consider a *linear* conditional variance function

$$\sigma_i^2 = E(e_i^2|\mathbf{z}_i) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} \tag{8.27}$$

Despite using a linear conditional variance function the test is for the general heteroskedasticity pattern in (8.24). Let $v_i = e_i^2 - E(e_i^2|\mathbf{z}_i)$ be the difference between a squared error and its conditional mean. Then, from (8.27), we can write

$$e_i^2 = E(e_i^2|\mathbf{z}_i) + v_i = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.28}$$

This looks very much like a linear regression model. The one problem is that the "dependent variable" e_i^2 is not observable. We overcome this problem by replacing e_i^2 with the squared OLS residuals $\hat{e}_i^2$. In large samples, this is valid because, as we show in Appendix 8B, the difference $e_i - \hat{e}_i$ goes to zero as $N \to \infty$. An operational version of (8.28) is

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8.29}$$

Strictly speaking, replacing e_i^2 by $\hat{e}_i^2$ also changes the definition of v_i, but we will retain the same notation to avoid unnecessary complication.

The test for heteroskedasticity is based on OLS estimation of (8.29). The question we ask is, do the variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ help explain $\hat{e}_i^2$? Under homoskedasticity the variables $z_{i2}, z_{i3}, \ldots, z_{iS}$ should have no relation to $\hat{e}_i^2$. One alternative is to use an F-test for the null hypothesis. An asymptotically equivalent and convenient test is based on the R^2, goodness-of-fit statistic, from (8.29). If the null hypothesis is true, $\alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$, then the R^2 should be small and close to zero. If R^2 is large, it is evidence against the assumption of homoskedasticity. How large does R^2 have to be for us to reject homoskedasticity? An answer requires a test statistic and a rejection region. It can be shown that if the random errors are homoskedastic, then the sample size multiplied by R^2, $N \times R^2$ or simply NR^2, has a chi-square (χ^2) distribution with $S - 1$ degrees of freedom in large samples. That is,

$$NR^2 \overset{a}{\sim} \chi^2_{(S-1)} \text{ if the null hypothesis of homoskedasticity is true} \tag{8.30}$$

Your exposure to the χ^2 distribution has been relatively limited. It is discussed in Appendix B.5.2. It was used for testing for normality in Section 4.3.4, and its relationship with the F-test was explored in Section 6.1.5. It is a distribution that is used for testing many different kinds of hypotheses. Like an F random variable, a χ^2 random variable only takes positive values. Critical values of the distribution appear in Statistical Table 3. Locate the test degrees of freedom in the left-hand column, and find the critical value from the columns, each of which corresponds to a percentile of the distribution. Because a large R^2 value is evidence against the null hypothesis of homoskedasticity (it suggests the z variables explain some changes in the variance), the rejection region for the statistic in (8.30) is in the right tail of the distribution. For an α-significance level test, we reject H_0 and conclude that heteroskedasticity exists when $NR^2 \geq \chi^2_{(1-\alpha,\, S-1)}$.

For example, if $\alpha = 0.01$ and $S = 2$, reject the hypothesis of homoskedasticity if $NR^2 \geq \chi^2_{(0.99,1)} = 6.635$. Your econometric software will have functions to calculate critical values, and p-values, for χ^2-tests.

There are several important features of this test:

1. It is a large sample test. The result in (8.30) holds approximately in large samples.
2. You will often see the test referred to as a **Lagrange multiplier test** (LM test) or a **Breusch–Pagan test** for heteroskedasticity. Breusch and Pagan used the LM principle (see Appendix C.8.4) to derive an earlier version of the test, which was later modified by other researchers to the form in (8.30). The test values for these and other slightly different versions of the test, one of which is the F-test, are automatically calculated by a number of software packages. The one provided by your software may or may not be exactly the same as the NR^2 version in (8.30). The relationships between the different versions of the test are described in Appendix 8B. As you proceed through the book and study more econometrics, you will find that many LM tests can be written in the form NR^2, where the R^2 comes from a convenient auxiliary regression related to the hypothesis being tested.
3. We motivated the test in terms of an alternative hypothesis with the very general conditional variance function $\sigma_i^2 = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS})$, yet we proceeded to carry out the test using the linear function $\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i$. One of the amazing features of the Breusch–Pagan/LM test is that the value of the statistic computed from the linear function is valid for testing an alternative hypothesis of heteroskedasticity where the variance function can be of any form given by (8.24).
4. The Breusch–Pagan test is for conditional heteroskedasticity. **Unconditional heteroskedasticity** exists when the error term variance is completely random, changing from observation to observation but unrelated to any particular variable. The least squares estimator properties are unaffected by unconditional heteroskedasticity. We illustrate this point in Appendix 8D.

8.6.4 The White Test

One problem with the variance function test described so far is that it presupposes that we have knowledge of what variables will appear in the variance function if the alternative hypothesis of heteroskedasticity is true. In other words, it assumes we are able to specify $z_2, z_3, \ldots, z_S$. In reality, we may wish to test for heteroskedasticity without precise knowledge of the relevant variables. With this point in mind, White suggested defining the z's as equal to the x's, the squares of the x's, and their cross-products. Frequently, the variables that affect the variance are the same as those in the mean function. Also, by using a quadratic function we can approximate a number of other possible conditional variance functions. Suppose the regression model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

The **White test** uses

$$z_2 = x_2 \quad z_3 = x_3 \quad z_4 = x_2^2 \quad z_5 = x_3^2 \text{ and } z_6 = x_2 x_3$$

If the regression model contains quadratic terms $(x_3 = x_2^2$ for example$)$, then some of the z's are redundant and are deleted. Also if x_3 is an indicator variable, taking the values 0 and 1, then $x_3^2 = x_3$ which is also redundant.

The White test is performed using the NR^2 test defined in (8.29), or an F-test (see Appendix 8B for details). One difficulty with the White test is that it can detect problems other than heteroskedasticity. Thus, while it is a useful diagnostic, be careful about interpreting the result of a significant White test. It may be that you have an incorrect functional form, or an omitted variable. In this sense, it is something like RESET, a specification error test discussed in Chapter 6.

8.6.5 Model Specification and Heteroskedasticity

As hinted at the end of the previous section, heteroskedasticity can be present because of a model specification error. If data partitions are not recognized, or important variables omitted, or an incorrect functional form selected, then heteroskedasticity can appear to be present. Hence, one piece of advice is to "Trust no one." Don't necessarily believe that a significant heteroskedasticity test means that heteroskedasticity is the problem and that using robust standard errors will be an adequate fix. Critically examine the model from the point of view of economic reasoning and look for any specification problems.

One very common specification issue with economic data is the choice of functional form. In Section 4.3.2, we discussed a variety of model specifications that are useful when considering nonlinear, or curvilinear, relationships (see Figure 4.5). Many economic applications use "log-log" or "log-linear" models. Using a logarithmic transformation of the dependent variable has another feature, **variance stabilization**, that is useful in the context of heteroskedastic data.[3] Economic variables like wages, incomes, house prices, and expenditures are right-skewed, with a long tail to the right. The **log-normal** probability distribution is useful when modeling such variables. This idea was introduced first in the Probability Primer in Figure P.2, and we discuss the log-normal distribution in Appendix B.3.9. If the random variable y has a log-normal probability density function, then $\ln(y)$ has a normal distribution, which is symmetrical and bell-shaped, and not skewed. That is, $\ln(y) \sim N(\mu, \sigma^2)$. The feature of the log-normal random variable that we are now interested in is that its variance increases when its mean and median increase. This is illustrated in Appendix B.3.9, Figure B.10, and the surrounding discussion. In Figure 8.6 we modify Figure 4.5(e) for the log-linear model to show $E(y|x)$, the solid line, and include $E(y|x) \pm 2\sqrt{\text{var}(y|x)}$, the dashed lines. By choosing a log-linear or log-log model we are implicitly assuming a curvilinear and heteroskedastic relationship between the variables y and x. However, there is a linear and homoskedastic relation between $\ln(y)$ and x.

Let's look at an example.

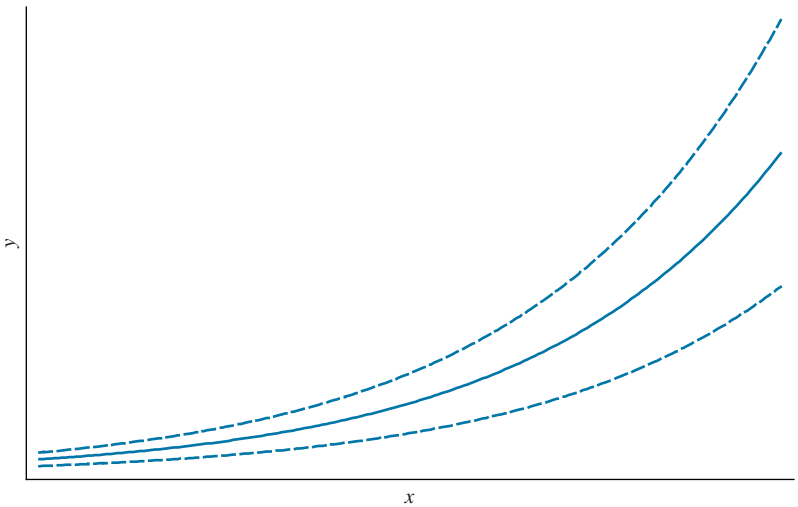


FIGURE 8.6 **A log-linear relationship.**

[3]The "Box-Cox Model" nests the linear and log-linear models in a more general nonlinear regression framework. See William Greene (2018) *Econometric Analysis, Eighth Edition*, 214–216.

EXAMPLE 8.8 | Variance Stabilizing Log-transformation

Consider the data file *cex5_small*. Figure 8.7(a) shows a histogram of household expenditures on entertainment per person, *ENTERT*, for those households who have positive spending, and Figure 8.7(b) is the histogram for ln(*ENTERT*).

Note the extremely skewed distribution of entertainment expenditures in Figure 8.7(a). Figure 8.7(b) shows the effect of the log-transformation. The distribution of ln(*ENTERT*) exhibits little skewness. Figure 8.8(a) shows the entertainment expenses plotted versus income and the least squares fitted line.

The variation in *ENTERT* about the fitted line increases as *INCOME* increases. Estimating the model $ENTERT = \beta_1 + \beta_2 INCOME + \beta_3 COLLEGE + \beta_4 ADVANCED + e$, we obtain the least squares residuals and then estimate by OLS the model $\hat{e}_i^2 = \alpha_1 + \alpha_2 INCOME_i + v_i$. From this regression, $NR^2 = 31.34$. The critical value for a 1% level of significance, heteroskedasticity test is 6.635, thus we conclude that heteroskedasticity is present. Figure 8.8(b) shows the log of entertainment expenses, ln(*ENTERT*), plotted versus income and the least squares fitted line. There is little if any visual evidence of heteroskedasticity and the value of the heteroskedasticity test statistic is $NR^2 = 0.36$, so we do not reject the null hypothesis of homoskedasticity. The log-transformation has "cured" the heteroskedasticity problem.

Among the 1200 households in the sample, 100 did not report any spending on entertainment. The log-transformation can only be used for positive values. We dropped the 100 with no spending, but that is not necessarily the best approach. In Section 16.7 we will discuss this type of data, which is called a **censored** sample.

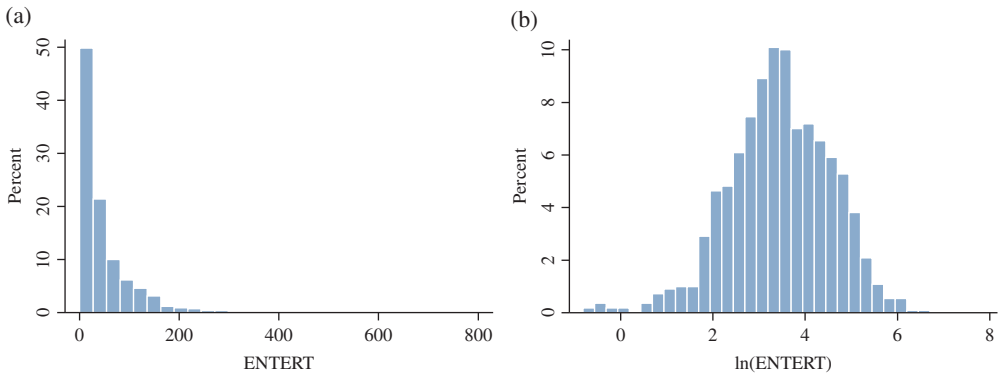


FIGURE 8.7 **Histograms of entertainment expenditures.**

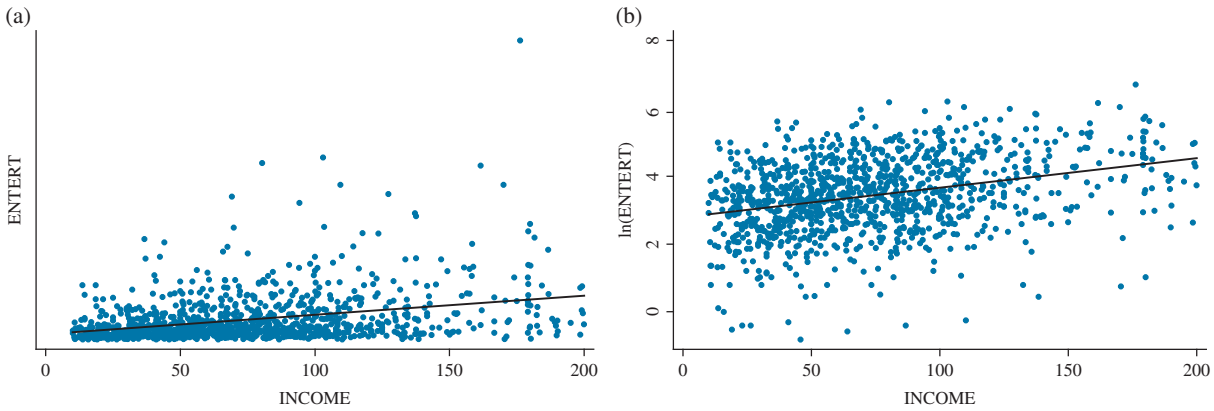


FIGURE 8.8 **Linear and log-linear models for entertainment expenditures.**

8.7 Heteroskedasticity in the Linear Probability Model

In Section 7.4 we introduced the **linear probability model** for explaining choice between two alternatives. We can represent this choice by an indicator variable y that takes the value one with probability p if the first alternative is chosen, and the value zero with probability $1 - p$ if the second alternative is chosen. An indicator variable with these properties is a Bernoulli random variable with mean $E(y) = p$ and variance $\text{var}(y) = p(1 - p)$. Interest centers on measuring the effect of explanatory variables $x_2, x_3, \ldots, x_k$ on the probability p. In the linear probability model the relationship between p and the explanatory variables is specified as the linear function

$$E(y_i|\mathbf{x}_i) = p = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

Defining the error e_i as the difference $y_i - E(y_i|\mathbf{x}_i)$ for the ith observation, we have the model

$$y_i = E(y_i|\mathbf{x}_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{8.31}$$

This model can be estimated with least squares—an example was given in Section 7.4—but it suffers from heteroskedasticity because

$$\begin{aligned} \text{var}(y_i|\mathbf{x}_i) &= \text{var}(e_i|\mathbf{x}_i) = p_i(1 - p_i) \\ &= (\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})(1 - \beta_1 - \beta_2 x_{i2} - \cdots - \beta_K x_{iK}) \end{aligned} \tag{8.32}$$

The error variance depends on the values of the explanatory variables. We can rectify this problem by applying the techniques described earlier in this chapter. Instead of using least squares standard errors, we can use heteroskedasticity-robust standard errors. Or, alternatively, we can apply a GLS procedure.

The first step toward obtaining GLS estimates is to estimate the variance in (8.32). An estimate of p_i can be obtained from the least squares predictions

$$\hat{p}_i = b_1 + b_2 x_{i2} + \cdots + b_K x_{iK} \tag{8.33}$$

giving an estimated variance of

$$\widehat{\text{var}}(e_i|\mathbf{x}) = \hat{p}_i(1 - \hat{p}_i) \tag{8.34}$$

A word of caution is required at this point. It is possible that some of the $\hat{p}_i$ obtained from (8.33) will not lie within the interval $0 < \hat{p}_i < 1$. If that happens, the corresponding variance estimate in (8.34) will be negative or zero, a nonsensical outcome. Thus, before proceeding to calculate the estimated variances from (8.34), it is necessary to check the estimated probabilities from (8.33) to ensure that they lie between zero and one. For those observations that violate this requirement, one possible solution is to set $\hat{p}_i$'s greater than 0.99 equal to 0.99, and $\hat{p}_i$'s less than 0.01 equal to 0.01. Another possible solution is to omit the offending observations. Neither of these solutions is totally satisfactory. Truncating at 0.99 or 0.01 is arbitrary, and the results could be sensitive to the truncation point. Omitting observations means that we are throwing away information. It might be preferable to use least squares with robust standard errors—that should, at least, be one of the options that is tried.

Once positive variance estimates have been obtained using (8.34), with adjustments where necessary, GLS estimates can be obtained by applying least squares to the transformed equation

$$\frac{y_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \beta_1 \frac{1}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \beta_2 \frac{x_{i2}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

EXAMPLE 8.9 | The Marketing Example Revisited

In Example 7.7 the choice of purchasing either Coke ($COKE = 1$) or Pepsi ($COKE = 0$) was modeled as depending on the relative price of Coke to Pepsi (*PRATIO*), and whether store displays for Coke and Pepsi were present (*DISP_COKE* = 1 if a Coke display was present, otherwise 0; *DISP_PEPSI* = 1 if a Pepsi display was present, otherwise 0). The data file *coke* contains 1140 observations on these variables. Table 8.2 contains the results for (1) least squares, (2) least squares with robust standard errors, (3) GLS with variances below 0.01 truncated to 0.01, and (4) GLS with observations not satisfying $0 < \hat{p}_i < 1$ omitted. For the GLS estimates there were no observations for which $\hat{p}_i > 0.99$ and there were 16 observations where $\hat{p}_i < 0.01$; for these latter cases it was also true that $\hat{p}_i < 0$.

Since the variance function in (8.32) contains the x's, their squares, and their cross products, a suitable test for heteroskedasticity is the White test described in Section 8.6.4. Applying this test to the residuals from the least squares estimated equation yields

$$\chi^2 = N \times R^2 = 25.817 \quad p\text{-value} = 0.0005$$

leading us to reject a null hypothesis of homoskedasticity at a 1% level of significance. Note that, when carrying out this test, your software will omit the squares of *DISP_COKE* and *DISP_PEPSI*. Because these variables are indicator variables, $DISP\_COKE^2 = DISP\_COKE$ and $DISP\_PEPSI^2 = DISP\_PEPSI$, leaving a χ^2-test with 7 degrees of freedom.

Examining the estimates in Table 8.2, we see there is little difference in the four sets of standard errors. In this particular case the use of least squares standard errors does not seem to matter. The four sets of coefficient estimates are also similar with the exception of those from GLS where the negative $\hat{p}$'s were truncated to 0.01. The weight on observations with variance $\text{var}(e_i) = 0.01(1 - 0.01) = 0.0099$ is a relatively large one. It appears that the large weights placed on those 16 observations are having a noticeable impact on the estimates. The signs are all as expected. Making Coke more expensive leads more people to purchase Pepsi. A Coke display encourages purchase of Coke, and a Pepsi display encourages purchase of Pepsi.

In Chapter 16 we study models which are specifically designed for modeling choice between two or more alternatives, and which do not suffer from the problems of the linear probability model.

TABLE 8.2 Linear Probability Model Estimates

	LS	LS-robust	GLS-trunc	GLS-omit
C	0.8902	0.8902	0.6505	0.8795
	(0.0655)	(0.0652)	(0.0568)	(0.0594)
PRATIO	−0.4009	−0.4009	−0.1652	−0.3859
	(0.0613)	(0.0603)	(0.0444)	(0.0527)
DISP	0.0772	0.0772	0.0940	0.0760
_COKE	(0.0344)	(0.0339)	(0.0399)	(0.0353)
DISP	−0.1657	−0.1657	−0.1314	−0.1587
_PEPSI	(0.0356)	(0.0343)	(0.0354)	(0.0360)

8.8 Exercises

8.8.1 Problems

8.1 For the simple regression model with heteroskedasticity, $y_i = \beta_1 + \beta_2 x_i + e_i$ and $\text{var}(e_i|\mathbf{x}_i) = \sigma_i^2$ show that the variance $\text{var}(b_2|\mathbf{x}_i) = \left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^{-1} \left[\sum_{i=1}^N (x_i - \bar{x})^2 \sigma_i^2\right] \left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^{-1}$ reduces to $\text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2$ under homoskedasticity.

8.2 Consider the regression model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ with two explanatory variables, x_{i1} and x_{i2}, but no constant term.

a. The sum of squares function is $S(\beta_1, \beta_2 | \mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{N} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$. Find the partial derivatives with respect to the parameters β_1 and β_2. Setting these derivatives to zero and solving, as in Appendix 2A, show that the least squares estimator of β_2 is

$$b_2 = \frac{\left(\sum x_{i1}^2\right)\left(\sum x_{i2} y_i\right) - \left(\sum x_{i1} x_{i2}\right)\left(\sum x_{i1} y_i\right)}{\left(\sum x_{i1}^2\right)\left(\sum x_{i2}^2\right) - \left(\sum x_{i1} x_{i2}\right)^2}$$

b. Let $x_{i1} = 1$ and show that the estimator in (a) reduces to

$$b_2 = \frac{\frac{\sum x_{i2} y_i}{N} - \frac{\sum x_{i2}}{N}\frac{\sum y_i}{N}}{\frac{\sum x_{i2}^2}{N} - \left(\frac{\sum x_{i2}}{N}\right)^2}$$

Compare this equation to equation (2A.5) and show that they are equivalent.

c. In the estimator in part (a), replace y_i, x_{i1} and x_{i2} by $y_i^* = y_i/\sqrt{h_i}$, $x_{i1}^* = x_{i1}/\sqrt{h_i}$ and $x_{i2}^* = x_{i2}/\sqrt{h_i}$. These are transformed variables for the heteroskedastic model $\sigma_i^2 = \sigma^2 h(\mathbf{z}_i) = \sigma^2 h_i$. Show that the resulting GLS estimator can be written as

$$\hat{\beta}_2 = \frac{\sum a_i x_{i2} y_i - \sum a_i x_{i2} \sum a_i y_i}{\sum a_i x_{i2}^2 - \left(\sum a_i x_{i2}\right)^2}$$

where $a_i = 1/(c h_i)$ and $c = \sum(1/h_i)$. Find $\sum_{i=1}^{N} a_i$.

d. Show that under homoskedasticity $\hat{\beta}_2 = b_2$.

e. Explain how $\hat{\beta}_2$ can be said to be constructed from "weighted data averages" while the usual least squares estimator b_2 is constructed from "arithmetic data averages." Relate your discussion to the difference between WLS and ordinary least squares.

8.3 Suppose that an outcome variable $y_{ij} = \beta_1 + \beta_2 x_{ij} + e_{ij}$, $i = 1, \ldots, N$; $j = 1, \ldots, N_i$. Assume $E(e_{ij}|\mathbf{X}) = 0$ and $\text{var}(e_{ij}|\mathbf{X}) = \sigma^2$. One illustration is y_{ij} = the ith farm's production on the jth acre of land, with each farm consisting of N_i acres. The variable x_{ij} is the amount of an input, labor or fertilizer, used by the ith farm on the jth acre.

a. Suppose that we do not have data on each individual acre, but only aggregate, farm-level data, $\sum_{j=1}^{N_i} y_{ij} = y_{Ai}$, $\sum_{j=1}^{N_i} x_{ij} = x_{Ai}$. If we specify the linear model $y_{Ai} = \beta_1 + \beta_2 x_{Ai} + e_{Ai}$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

b. Suppose that we do not have data on each individual acre, but only average data for each farm, $\sum_{j=1}^{N_i} y_{ij}/N_i = \bar{y}_i$, $\sum_{j=1}^{N_i} x_{ij}/N_i = \bar{x}_i$. If we specify the linear model $\bar{y}_i = \beta_1 + \beta_2 \bar{x}_i + \bar{e}_i$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

c. Suppose the outcome variable is binary. For example, suppose $y_{ij} = 1$ if a crop shows evidence of blight on the jth acre of the ith farm, and $y_{ij} = 0$ otherwise. In this case $\sum_{j=1}^{N_i} y_{ij}/N_i = p_i$, where p_i is the sample proportion of acres that show the blight on the ith farm. Suppose the probability of the ith farm showing blight on a particular acre is P_i. If we specify the linear model $\bar{y}_i = \beta_1 + \beta_2 \bar{x}_i + \bar{e}_i$, $i = 1, \ldots, N$, what is the conditional variance of the random error?

8.4 Consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where we hypothesize heteroskedasticity of the form $\sigma_i^2 = \sigma^2 x_i^2$. We have $N = 4$ observations, with $x = (1\ \ 2\ \ 3\ \ 4)$ and $y = (3\ \ 4\ \ 3\ \ 5)$.

a. Use the formula for the least squares estimator in Exercise 8.2(b) to compute the OLS estimate of β_2. In this case $\sum x_{i2} y_i / N = 10$, $\sum x_{i2}^2 / N = 7$.

b. Referring to Exercise 8.2(c), what is the value $c = \sum(1/h_i)$?

c. Referring to Exercise 8.2(c), what are the values $a_i = 1/(c h_i)$, $i = 1, \ldots, 4$? What is $\sum_{i=1}^{4} a_i$?

d. Use the formula for the generalized least squares estimator in Exercise 8.2(c) to compute the GLS estimate of β_2.

e. Suppose that we know that $\sigma^2 = 0.2$. Calculate the true OLS variance given in equation (8.8). The values of $(x_i - \bar{x})^2$ are $(2.25,\ 0.25,\ 0.25,\ 2.25)$. What is the value of the incorrect variance in equation (8.6)?

8.5 Consider the simple regression model $y_i = \beta_1 + \beta_2 x_{i2} + e_i$. Suppose $N = 5$ and the values of x_{i2} are $(1, 2, 3, 4, 5)$. Let the true values of the parameters be $\beta_1 = 1$, $\beta_2 = 1$. Let the true random error values, which are never known in reality, be $e_i = (1,\ -1,\ 0,\ 6,\ -6)$.

a. Calculate the values of y_i.

b. The OLS estimates of the parameters are $b_1 = 3.1$ and $b_2 = 0.3$. Compute the least squares residual, $\hat{e}_1$, for the first observation, and $\hat{e}_4$, for the fourth observation. What is the sum of all the least squares residuals? In this example, what is the sum of the true random errors? Is the sum of the residuals always equal to the sum of the random errors? Explain.

c. It is hypothesized that the data are heteroskedastic with the variance of the first three random errors being σ_1^2, and the variance of the last two random errors being σ_2^2. We regress the squared residuals $\hat{e}_i^2$ on the indicator variable z_i, where $z_i = 0,\ i = 1, 2, 3$ and $z_i = 1,\ i = 4, 5$. The overall model F-statistic value is 12.86. Does this value provide evidence of heteroskedasticity at the 5% level of significance? What is the p-value for this F-value (requires computer)?

d. $R^2 = 0.8108$ from the regression in (c). Use this value to carry out the LM (Breusch–Pagan) test for heteroskedasticity at the 5% level of significance. What is the p-value for this test (requires computer)?

e. We now regress $\ln\left(\hat{e}_i^2\right)$ on z_i. The estimated coefficient of z_i is 3.81. We discover that the software reports using only $N = 4$ observations in this calculation. Why?

f. In order to carry out feasible generalized least squares using information from the regression in part (e), we first create the transformed variables $\left(y_i^*,\ x_{i1}^*,\ x_{i2}^*\right)$. List the values of the transformed observations for $i = 1$ and $i = 4$.

8.6 Consider the wage equation

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i \qquad \text{(XR8.6a)}$$

where wage is measured in dollars per hour, education and experience are in years, and $METRO = 1$ if the person lives in a metropolitan area. We have $N = 1000$ observations from 2013.

a. We are curious whether holding education, experience, and *METRO* constant, there is the same amount of random variation in wages for males and females. Suppose $\text{var}\left(e_i|\mathbf{x}_i, FEMALE = 0\right) = \sigma_M^2$ and $\text{var}\left(e_i|\mathbf{x}_i, FEMALE = 1\right) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Using 577 observations on males, we obtain the sum of squared OLS residuals, $SSE_M = 97161.9174$. The regression using data on females yields $\hat{\sigma}_F = 12.024$. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. We hypothesize that married individuals, relying on spousal support, can seek wider employment types and hence holding all else equal should have more variable wages. Suppose $\text{var}\left(e_i|\mathbf{x}_i, MARRIED = 0\right) = \sigma_{SINGLE}^2$ and $\text{var}\left(e_i|\mathbf{x}_i, MARRIED = 1\right) = \sigma_{MARRIED}^2$. Specify the null hypothesis $\sigma_{SINGLE}^2 = \sigma_{MARRIED}^2$ versus the alternative hypothesis $\sigma_{MARRIED}^2 > \sigma_{SINGLE}^2$. We add *FEMALE* to the wage equation as an explanatory variable, so that

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + \beta_5 FEMALE + e_i \qquad \text{(XR8.6b)}$$

Using $N = 400$ observations on single individuals, OLS estimation of (XR8.6b) yields a sum of squared residuals is 56231.0382. For the 600 married individuals, the sum of squared errors is 100,703.0471. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

c. Following the regression in part (b), we carry out the NR^2 test using the right-hand-side variables in (XR8.6b) as candidates related to the heteroskedasticity. The value of this statistic is 59.03. What do we conclude about heteroskedasticity, at the 5% level? Does this provide evidence about the issue discussed in part (b), whether the error variation is different for married and unmarried individuals? Explain.

d. Following the regression in part (b) we carry out the White test for heteroskedasticity. The value of the test statistic is 78.82. What are the degrees of freedom of the test statistic? What is the 5% critical value for the test? What do you conclude?

e. The OLS fitted model from part (b), with usual and robust standard errors, is

$$\widehat{\text{WAGE}} = -17.77 + 2.50EDUC + 0.23EXPER + 3.23METRO - 4.20FEMALE$$
$$\begin{array}{lllll} \text{(se)} & (2.36)\ (0.14) & (0.031) & (1.05) & (0.81) \\ \text{(robse)} & (2.50)\ (0.16) & (0.029) & (0.84) & (0.80) \end{array}$$

For which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

f. If we add *MARRIED* to the model in part (b), we find that its t-value using a White heteroskedasticity robust standard error is about 1.0. Does this conflict with, or is it compatible with, the result in (b) concerning heteroskedasticity? Explain.

8.7 Consider the simple treatment effect model $y_i = \beta_1 + \beta_2 d_i + e_i$. Suppose that $d_i = 1$ or $d_i = 0$ indicating that a treatment is given to randomly selected individuals or not. The dependent variable y_i is the outcome variable. See the discussion of the difference estimator in Section 7.5.1. Suppose that N_1 individuals are given the treatment and N_0 individual are in the control group, who are not given the treatment. Let $N = N_0 + N_1$ be the total number of observations.

a. Show that if $\text{var}(e_i|\mathbf{d}) = \sigma^2$ then the variance of the OLS estimator b_2 of β_2 is $\text{var}(b_2|\mathbf{d}) = N\sigma^2/(N_0N_1)$. [*Hint*: See Appendix 7B.]

b. Let $\bar{y}_0 = \sum_{i=1}^{N_0} y_i/N_0$ be the sample mean of the outcomes for the N_0 observations on the control group. Let $SST_0 = \sum_{i=1}^{N_0}(y_i - \bar{y}_0)^2$ be the sum of squares about the sample mean of the control group, where $d_i = 0$. Similarly, let $\bar{y}_1 = \sum_{i=1}^{N_1} y_i/N_1$ be the sample mean of the outcomes for the N_1 observations on the treated group, where $d_i = 1$. Let $SST_1 = \sum_{i=1}^{N_1}(y_i - \bar{y}_1)^2$ be the sum of squares about the sample mean of the treatment group. Show that $\hat{\sigma}^2 = \sum_{i=1}^{N} \hat{e}_i^2/(N-2) = (SST_0 + SST_1)/(N-2)$ and therefore that

$$\widehat{\text{var}}(b_2|\mathbf{d}) = N\hat{\sigma}^2/(N_0N_1) = \left(\frac{N}{N-2}\right)\left(\frac{SST_0 + SST_1}{N_0N_1}\right)$$

c. Using equation (2.14) find $\text{var}(b_1|\mathbf{d})$, where b_1 is the OLS estimator of the intercept parameter β_1. What is $\widehat{\text{var}}(b_1|\mathbf{d})$?

d. Suppose that the treatment and control groups have not only potentially different means but potentially different variances, so that $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. Find $\text{var}(b_2|\mathbf{d})$. What is the unbiased estimator for $\text{var}(b_2|\mathbf{d})$? [*Hint*: See Appendix C.4.1.]

e. Show that the White heteroskedasticity robust estimator in equation (8.9) reduces in this case to $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Compare this estimator to the unbiased estimator in part (d).

f. What does the robust estimator become if we drop the degrees of freedom correction $N/(N-2)$ in the estimator proposed in part (e)? Compare this estimator to the unbiased estimator in part (d).

8.8 It can be shown that the theoretically useful form of the OLS estimator of β_1 in the simple linear regression model $y_i = \beta_1 + \beta_2 x_{i2} + e_i$ is $b_1 = \beta_1 + \sum(-\bar{x}w_i + N^{-1})e_i = \sum v_i e_i$, where $v_i = (-\bar{x}w_i + N^{-1})$ and $w_i = (x_i - \bar{x})/\sum(x_i - \bar{x})^2$. Using this formula consider the simple treatment effect model $y_i = \beta_1 + \beta_2 d_i + e_i$. Suppose that $d_i = 1$ or $d_i = 0$ indicating that a treatment is given to a randomly selected individual or not. The dependent variable y_i is the outcome variable. See the discussion of the difference estimator in Section 7.5.1. Suppose that N_1 individuals are given the treatment and N_0 individuals in the control group are not given the treatment. Let $N = N_0 + N_1$ be the total number of observations.

a. Show that when $d_i = 0$, $v_i = 1/N$ and that when $d_i = 1$, $v_i = 0$.

b. Derive $\text{var}(b_1|\mathbf{d})$ under the assumption of homoskedastic errors, $\text{var}(e_i|\mathbf{d}) = \sigma^2$. What is an unbiased estimator of $\text{var}(b_1|\mathbf{d})$ in this case?

c. Derive $\text{var}(b_1|\mathbf{d})$ under the assumption of heteroskedastic errors, $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. What is an unbiased estimator of $\text{var}(b_1|\mathbf{d})$ in this case?

8.9 We wish to estimate the hedonic regression model

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 CLOSE_i + \beta_4 AGE_i + \beta_5 FIREPLACE_i + \beta_6 POOL_i + \beta_7 TWOSTORY_i + e_i$$

The variables are *PRICE* ($1000), *SQFT* (100s), *CLOSE* = 1 if located near a major university, 0 otherwise, *AGE* (years), *FIREPLACE*, *POOL*, *TWOSTORY* = 1 if present, 0 otherwise.

a. Using Table 8.3, comment on the sign, significance, and interpretation of the OLS coefficient estimate for the variable *CLOSE*.

b. Answer each of the following True or False. In a regression model with heteroskedasticity, (i) the OLS estimator is biased; (ii) the OLS estimator is inconsistent; (iii) the OLS estimator does not have an approximate normal distribution in large samples; (iv) the usual OLS standard error is too small; (v) the usual OLS estimator standard error is incorrect; (vi) the usual R^2 is no longer meaningful; (vii) the usual overall F-test is reliable in large samples.

c. Following the OLS regression, the residuals are saved as *EHAT*. In the regression labeled AUX in Table 8.3, the dependent variable is $EHAT^2$. Test for the presence of heteroskedasticity, using the 5% level of significance. State the test statistic, the test critical value, and your conclusion.

d. The model is reestimated by OLS using White heteroskedasticity-consistent standard errors. In what way are these standard errors robust? Are they valid when there is homoskedasticity, heteroskedasticity, in small samples and large? Which of the statistically significant coefficients has wider confidence intervals using the robust standard errors? Do any coefficients switch from being significant at 5% to not significant at 5%, or vice versa?

e. Our researcher estimates the equation after dividing each variable, and the constant term, by *SQFT* to obtain the GLS estimates. What assumption has been made about the form of heteroskedasticity in this estimation? Are the GLS estimates, shown in Table 8.3, noticeably different from the OLS estimates? Do any coefficients switch from being significant at 5% to not significant at 5%, or vice versa?

f. The residuals from the transformed regression in part (e) are called *ESTAR*. The researcher regresses $ESTAR^2$ on all the transformed variables and includes an intercept. The $R^2 = 0.0237$. Has the researcher eliminated heteroskedasticity?

g. The researcher estimates the model in (e) again but uses robust standard errors. These are reported in Table 8.3 as "Robust GLS." Do you consider this a prudent thing to do? Explain your reasoning.

TABLE 8.3 **Estimates for Exercise 8.9**

	OLS	AUX	Robust OLS	GLS	Robust GLS
C	−101.072***	−25561.243***	−101.072***	−4.764	−4.764
	(27.9055)	(5419.9443)	(34.9048)	(21.1357)	(35.8375)
SQFT	13.3417***	1366.8074***	13.3417***	7.5803***	7.5803***
	(0.5371)	(104.3092)	(1.1212)	(0.5201)	(0.9799)
CLOSE	26.6657***	1097.8933	26.6657***	39.1988***	39.1988***
	(9.8602)	(1915.0902)	(9.6876)	(7.0438)	(7.2205)
AGE	−2.7305	52.4499	−2.7305	1.4887	1.4887
	(2.7197)	(528.2353)	(3.2713)	(2.1034)	(2.5138)
FIREPLACE	−2.2585	−3005.1375	−2.2585	17.3827**	17.3827*
	(10.5672)	(2052.4109)	(10.6369)	(7.9023)	(9.3531)
POOL	0.3601	6878.0158*	0.3601	8.0265	8.0265
	(19.1855)	(3726.2941)	(27.2499)	(17.3198)	(15.6418)
TWOSTORY	5.8833	−7394.3869**	5.8833	26.7224*	26.7224*
	(14.8348)	(2881.2790)	(20.8733)	(13.7616)	(16.0651)
R^2	0.6472	0.3028	0.6472	0.4427	0.4427

Standard errors in parentheses
*$p < 0.10$
**$p < 0.05$
***$p < 0.01$

8.10 Does having more children drive parents to drink more alcohol? We have data on the following variables: $WALC$ = budget share (percent of income spent) for alcohol expenditure; $INCOME$ = total net household income (10,000 UK pounds); AGE = age of household head/10; NK = number of children (1 or 2). We are interested in the equation

$$\ln(WALC) = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 NK + e$$

a. The data we have is based on a survey. If we hope to establish a causal relationship between NK and the budget share spent on alcohol, what assumptions are sufficient to prove that the least squares estimator is BLUE?

b. Using 1278 observations on households with a positive budget share for alcohol, the OLS estimated equation, with conventional standard errors, is

$$\widehat{\ln(WALC)} = -1.956 + 0.837 INCOME - 0.228 AGE - 0.251 NK$$
$$\text{(se)} \qquad (0.166) \quad (0.516) \qquad (0.039) \qquad (0.058)$$

Test the null hypothesis that an increase in the number of children from one to two has no effect on the budget share of alcohol versus the alternative that an increase in the number of children increases the budget share of alcohol. Use the 5% level of significance.

c. We suspect that the regression error variance might be larger for households with two children rather than one. We estimate the budget share equation by least squares separately for households with one and two children. For the 489 households with one child, the sum of squared residuals is 465.83. For the 789 households with two children, the sum of squared residuals is 832.77. Test the null hypothesis that there is no difference between the regression error variances for these two groups, against the alternative that there is a difference. Use the Goldfeld–Quandt test at the 5% level of significance. Repeat the test using the alternative that the regression error variance for the subsample of households with two children is greater than the regression error variance for the subsample of households with one child. What do you conclude?

d. We save the least squares residuals from the estimation in part (b), calling them $EHAT$. We then obtain the second-stage regression results $EHAT^2 = 0.012 + 0.279 AGE + 0.025 NK$ with an $R^2 = 0.0208$. Is there evidence of heteroskedasticity? Set up the appropriate hypothesis and carry out the test at the 1% level of significance. What do you conclude?

e. We then carry out the regression $\widehat{\ln(EHAT^2)} = -2.088 + 0.291 AGE - 0.048 NK$. Holding NK constant, calculate the estimated variance ratio $\widehat{\text{var}}(e_i|AGE = 40)/\widehat{\text{var}}(e_i|AGE = 30)$. [*Hint*: Recall that AGE is measured in units of 10 years.] What is the estimated ratio $\widehat{\text{var}}(e_i|AGE = 60)/\widehat{\text{var}}(e_i|AGE = 30)$? Holding AGE constant, calculate the estimated variance ratio $\widehat{\text{var}}(e_i|NK = 2)/\widehat{\text{var}}(e_i|NK = 1)$.

f. Based on the results we have obtained so far, can we claim that the least squares estimator used in (b) is BLUE?

g. What model would we estimate by OLS to implement feasible generalized least squares estimation?

8.11 We are interested in the relationship between rice production, inputs of labor and fertilizer, and the area planted using data on $N = 44$ farms.

$$RICE_i = \beta_1 + \beta_2 LABOR_i + \beta_3 FERT_i + \beta_4 ACRES_i + e_i$$

a. We observe the least squares residuals, $\hat{e}_i$, increase in magnitude when plotted against $ACRES$. We regress $\hat{e}_i^2$ on $ACRES$ and obtain a regression with $R^2 = 0.2068$. The estimated coefficient of $ACRES$ is 2.024 with the standard error of 0.612. What can we conclude about heteroskedasticity based on these results? Explain your reasoning.

b. We instead estimate the model

$$RICE_i/ACRES_i = \alpha + \beta_1(1/ACRES_i) + \beta_2 LABOR_i/ACRES_i + \beta_3 FERT_i/ACRES_i + e_i$$

What is the implicit assumption about the heteroskedasticity pattern?

c. Many economists would omit $(1/ACRES_i)$ from the equation. What argument can you propose that would make this defensible?

d. Following the estimation of the model in (b) or (c), the squared residuals, $\tilde{e}_i^2$, are regressed on $ACRES$. The estimated coefficient is negative and significant at the 10% level. The regression

$R^2 = 0.0767$. What might you conclude about the models in (b) or (c)? That is, what could have led to such results?

e. In a further step, we estimate $\ln\left(\hat{e}_i^2\right) = -1.30 + 1.11\ln(ACRES)$ and $\ln\left(\tilde{e}_i^2\right) = -1.20 - 1.21\ln(ACRES)$. What evidence does this provide about the question in part (d)?

f. If we estimate the model in (c), omitting $\left(1/ACRES_i\right)$, would you advise using White heteroskedasticity robust standard errors? Explain why or why not.

8.12 An econometrician wishes to study the properties of an estimator using simulated data. Suppose the sample size N is set to be 100. The intercept and slope parameters are 100, and 10, respectively. The one explanatory variable, x, has a normal distribution with mean 10 and standard deviation 10. A standard normal random variable, z, independent of x, is created. The data generating process is $y_i = \beta_1 + \beta_2 x_i + e_i$, where

$$e_i = \begin{cases} z_i & \text{if } i \text{ is an odd number} \\ 2z_i & \text{if } i \text{ is an even number} \end{cases}$$

a. The OLS estimator is not the best linear unbiased estimator using the 100 data pairs $\left(y_i, x_i\right)$. True or false? Explain.

b. If we divide y and x for the even number observations by $\sqrt{2}$, leaving the odd number observations alone, and then run a least squares regression, the resulting estimator is BLUE. True or false? Explain.

c. Suppose you were assigned the task of showing that the heteroskedasticity in the data was "statistically significant." Using the 100 data pairs $\left(y_i, x_i\right)$, how exactly would you do it?

8.13 A researcher has 1100 observations on household expenditures on entertainment (per person in the previous quarter, $) *ENTERT*. The researcher wants to explain these expenditures as a function of *INCOME* (monthly income during past year, $100 units), whether the household lives in an *URBAN* area, and whether someone in the household has a *COLLEGE* degree (Bachelor's) or an *ADVANCED* degree (Master's or Ph.D.). *COLLEGE* and *ADVANCED* are indicator variables.

a. The OLS estimates and t-values are given in Table 8.4, on the next page. Taking the residuals from this regression, and regressing their squared values on all explanatory variables yields an $R^2 = 0.0344$. Such a small value implies there is no heteroskedasticity, correct? If that statement is not correct, then carry out the proper test. What do you conclude about the presence of heteroskedasticity?

b. To be safe the researcher uses White heteroskedasticity robust standard errors, given in Table 8.4. The researcher's paper has to do with the effect on entertainment expenditures of having someone with an advanced degree in the household. Compare the significance of *ADVANCED* in the two OLS regressions. What do you find? It is generally true that robust standard errors are larger than ones that are not robust. Is that true or false in this case?

c. Because of the importance of the variable *ADVANCED* in the model, the researcher takes some additional effort. Using the OLS residuals $\hat{e}_i$, the researcher obtains

$$\ln\left(\hat{e}_i^2\right) = 4.9904 + 0.0177INCOME_i + 0.2902ADVANCED_i$$
$$(t) \qquad (10.92) \qquad (1.80)$$

What evidence about heteroskedasticity is present in these results?

d. The researcher takes the results in (c) and then calculates

$$h_i = \exp\left(0.0177INCOME_i + 0.2902ADVANCED_i\right)$$

Each variable, including the intercept, is divided by $\sqrt{h_i}$ and the model reestimated to obtain the FGLS results in Table 8.4. Based on these results, how much of an effect on entertainment expenditures is there for households including someone with an advanced degree? Is this statistically significant? To which set of OLS results, can we make a valid comparison with the FGLS estimates? Have we improved the estimation of the effect of *ADVANCED* on entertainment by taking the steps in (c) and (d)? Provide a very careful answer to this question.

e. Looking for an easier way the researcher estimates a log-linear model shown in Table 8.4. Following this estimation, regressing the squared residuals on the explanatory variables, we find $NR^2 = 2.46$. Using White's test, including all the squares and cross-products of the explanatory variables, we obtain $NR^2 = 6.63$. What are the critical values for each of these test statistics? Using a test at the 5% level, do we reject homoskedasticity in the log-linear model or not?

f. Interpret the regression results in (e) from the point of view of the researcher interested in the effect of *ADVANCED* on entertainment expenditures. What exactly has happened by using the log-linear model? Provide an intuitive explanation. As a hint, Figure 8.9 shows entertainment expenditures for one range of income, between $7000/mo and $8000/mo.

TABLE 8.4 Estimates for Exercise 8.13

	OLS	Robust OLS	FGLS	Log-linear
C	20.5502	20.5502	18.5710	2.7600
	(3.19)	(3.30)	(4.16)	(25.79)
INCOME	0.5032	0.5032	0.4447	0.0080
	(10.17)	(6.45)	(8.75)	(9.77)
URBAN	−6.4629	−6.4629)	−0.8420	0.0145
	(−1.06)	(−0.81)	(−0.20)	(0.14)
COLLEGE	−0.7155	−0.7155	1.7388	0.0576
	(−0.16)	(−0.15)	(0.52)	(0.77)
ADVANCED	9.8173	9.8173	9.0123	0.2315
	(1.87)	(1.58)	(1.92)	(2.65)

t-values in parentheses

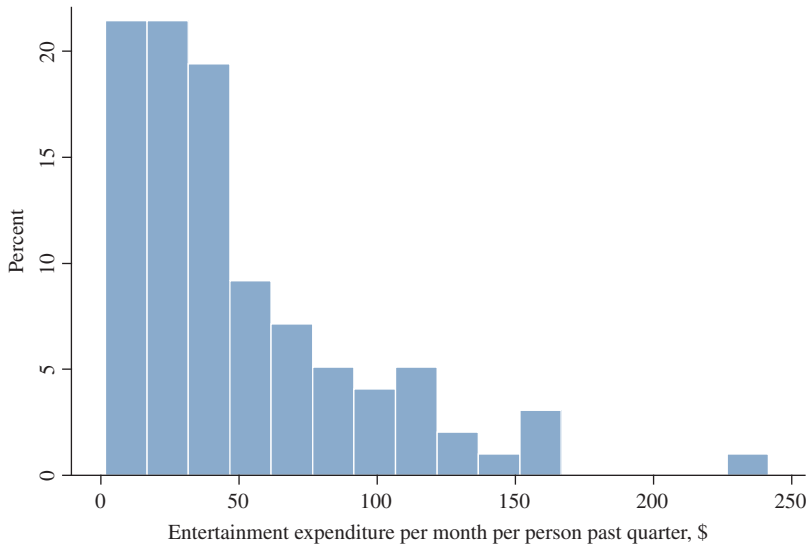


FIGURE 8.9 Histogram for entertainment expenditure.

8.14 Using data on 1000 home loan borrowers, we estimate the linear probability model

$$DEFAULT = \beta_1 + \beta_2 LTV + \beta_3 RATE + \beta_4 AMOUNT + \beta_5 FICO + e$$

where $DEFAULT = 1$ if the borrower has made a mortgage payment more than 90 days late, $LTV = 100$(loan amount/property value), *RATE* is the interest rate, *AMOUNT* ($10,000 units) of the loan, and *FICO* is the borrower's credit score.

a. Figure 8.10(a) is the histogram of the least squares residuals, $\hat{e}$. Explain the bimodal shape.

b. Figure 8.10(b) is the histogram of the least squares fitted values,

$$\widehat{DEFAULT} = 0.6887 + 0.0055LTV + 0.0482RATE - 0.0012AMOUNT - 0.0014FICO$$

Explain the interpretation of the fitted values. Do you find any unusual fitted values in the figure?

c. Let Y be a Bernoulli random variable, taking the values 1 and 0 with probabilities P and $1 - P$. Show that $\text{var}(Y) = P(1 - P)$.

d. Regressing $\hat{e}_i^2$ on the explanatory variables, we obtain $R^2 = 0.0206$ and the model F-statistic is 5.22. What does each of these values tell us about the null hypothesis of homoskedasticity in this model? Provide any relevant test statistics, and their 5% level of significance critical values. In light of part (c), are the results surprising?

e. Consider two hypothetical borrowers:

Borrower 1: $LTV = 85$, $RATE = 11$, $AMOUNT = 400$, $FICO = 500$

Borrower 2: $LTV = 50$, $RATE = 5$, $AMOUNT = 100$, $FICO = 700$

The 95% interval estimates, for the expected probability of default for the hypothetical borrowers using OLS, OLS with heteroskedasticity robust standard errors, and FGLS are given in Table 8.5. Discuss these interval estimates. If two such borrowers came for a loan, to whom would you offer one?

f. To obtain the FGLS estimates in (e), negative predicted values in nine observations are turned to positives by taking their absolute value. Why did we do that? What other alternatives did we have?

TABLE 8.5 **Interval Estimates for Exercise 8.14(e)**

Borrower	Method	Lower Bound	$\widehat{DEFAULT}$	Upper Bound	Std. Err.
1	OLS	−0.202	0.527	1.257	0.372
1	OLS (robust)	−0.132	0.527	1.187	0.337
1	FGLS	−0.195	0.375	0.946	0.291
2	OLS	−0.043	0.116	0.275	0.082
2	OLS (robust)	−0.025	0.116	0.257	0.072
2	FGLS	−0.019	0.098	0.215	0.060

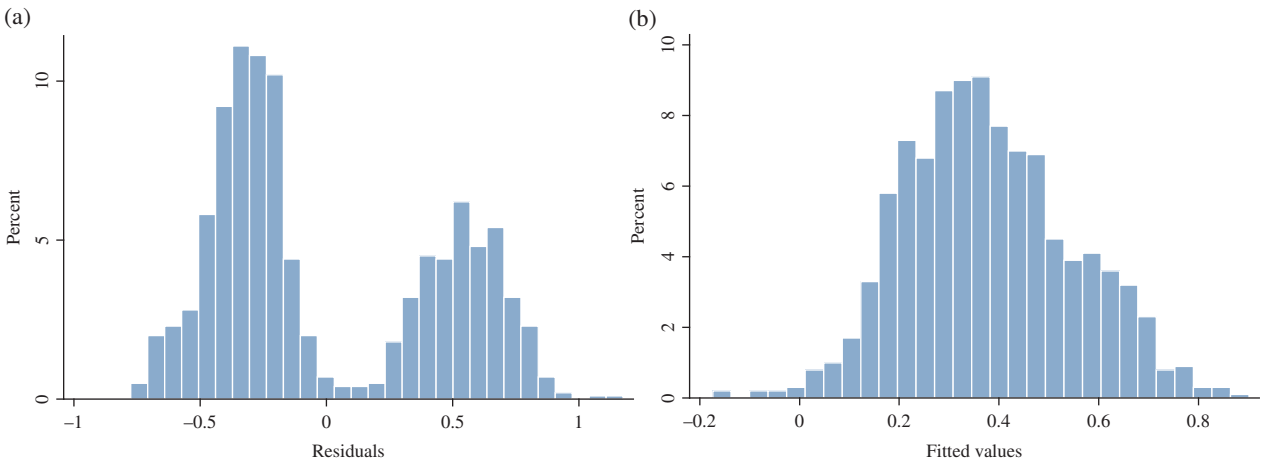


FIGURE 8.10 **Histograms for residuals and fitted values for Exercise 8.14.**

8.15 We have $N = 396$ observations on employment at fast-food restaurants in two neighboring states, New Jersey and Pennsylvania. In Pennsylvania, the control group $d_i = 0$, there is no minimum wage law. In New Jersey, the treatment group $d_i = 1$, there is a minimum wage law. Let the observed outcome variable be full-time employment FTE_i at comparable fast-food restaurants. Some sample summary statistics for FTE_i in the two states are in Table 8.6. For Pennsylvania, the sample size is $N_0 = 77$, the sample mean is $\overline{FTE}_0 = \sum_{i=1,d_i=0}^{N_0} FTE_i/N_0$, the sample variance is $s_0^2 = \sum_{i=1,d_i=0}^{N_0}\left(FTE_i - \overline{FTE}_0\right)^2/(N_0 - 1) = SST_0/(N_0 - 1)$, the sample standard deviation is $s_0 = \sqrt{s_0^2}$, and the standard error of mean is $se_0 = \sqrt{s_0^2/N_0} = s_0/\sqrt{N_0}$. For New Jersey, the definitions are comparable with subscripts "1".

TABLE 8.6 **Summary Statistics for Exercise 8.15**

State	d	N	Sample Mean	Sample Variance	Sample Standard Deviation	Standard Error of the Mean
Pennsylvania (control)	0	77	21.16558	68.50429	8.276732	0.9432212
New Jersey (treatment)	1	319	21.02743	86.36029	9.293024	0.5203094

a. Consider the regression model $FTE_i = \beta_1 + \beta_2 d_i + e_i$. The OLS estimates are given below, along with the usual standard errors (se), the White heteroskedasticity robust standard errors (robse), and an alternative robust standard error (rob2).

$$\begin{aligned} \widehat{FTE}_i &= 21.16558 - 0.1381549 d_i \\ \text{(se)} \quad & (1.037705) \quad (1.156182) \\ \text{(robse)} \quad & (0.9394517) \quad (1.074157) \\ \text{(rob2)} \quad & (0.9432212) \quad (1.077213) \end{aligned}$$

Show the relationship between the least squares estimates of the coefficients, the estimated slope and intercept, and the summary statistics in Table 8.6.

b. Calculate $\widehat{\text{var}}(b_2|\mathbf{d}) = N\hat{\sigma}^2/(N_0 N_1) = \left(\frac{N}{N-2}\right)\left(\frac{SST_0 + SST_1}{N_0 N_1}\right)$, derived in Exercise 8.7(b). Compare the standard error of the slope using this expression to the regression output in part (a).

c. Suppose that the treatment and control groups have not only potentially different means but potentially different variances, so that $\text{var}(e_i|d_i = 1) = \sigma_1^2$ and $\text{var}(e_i|d_i = 0) = \sigma_0^2$. Carry out the Goldfeld–Quandt test of the null hypothesis $\sigma_0^2 = \sigma_1^2$ at the 1% level of significance. [*Hint*: See Appendix C.7.3.]

d. In Exercise 8.7(e), we showed that the heteroskedasticity robust variance for the slope estimator is $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Use the summary statistic data to calculate this quantity. Compare the heteroskedasticity robust standard error of the slope using this expression to those from the regression output. In Appendix 8D, we discuss several heteroskedasticity robust variance estimators. This one is most common and usually referred to as "HCE1," where HCE stands for "heteroskedasticity consistent estimator."

e. Show that the alternative robust standard error, rob2, can be computed from $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{SST_0}{N_0(N_0-1)} + \frac{SST_1}{N_1(N_1-1)}$. In Appendix 8D, this estimator is called "HCE2." Note that it can be written $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\hat{\sigma}_0^2/N_0\right) + \left(\hat{\sigma}_1^2/N_1\right)$, where $\hat{\sigma}_0^2 = SST_0/(N_0 - 1)$ and $\hat{\sigma}_1^2 = SST_1/(N_1 - 1)$. These estimators are unbiased and are discussed in Appendix C.4.1. Is the variance estimator unbiased if $\sigma_0^2 = \sigma_1^2$?

f. The estimator HCE1 is $\widehat{\text{var}}(b_2|\mathbf{d}) = \frac{N}{N-2}\left(\frac{SST_0}{N_0^2} + \frac{SST_1}{N_1^2}\right)$. Show that dropping the degrees of freedom correction $N/(N-2)$ it becomes HCE0, $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\tilde{\sigma}_0^2/N_0\right) + \left(\tilde{\sigma}_1^2/N_1\right)$, where $\tilde{\sigma}_0^2 = SST_0/N_0$ and $\tilde{\sigma}_1^2 = SST_1/N_1$ are biased but consistent estimators of the variances. See Appendix C.4.2. Calculate the standard error for b_2 using this alternative.

g. A third variant of a robust variance estimator, HCE3, is $\widehat{\text{var}}(b_2|\mathbf{d}) = \left(\frac{\hat{\sigma}_0^2}{N_0 - 1}\right) + \left(\frac{\hat{\sigma}_1^2}{N_1 - 1}\right)$, where $\hat{\sigma}_0^2 = SST_0/(N_0 - 1)$ and $\hat{\sigma}_1^2 = SST_1/(N_1 - 1)$. Calculate the robust standard error using HCE3 for this example. In this application, comparing HCE0 to HCE2 to HCE3, which is largest? Which is smallest?

8.8.2 Computer Exercises

8.16 A sample of 200 Chicago households was taken to investigate how far American households tend to travel when they take a vacation. Consider the model

$$MILES = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 KIDS + e$$

MILES is miles driven per year, *INCOME* is measured in \$1000 units, *AGE* is the average age of the adult members of the household, and *KIDS* is the number of children.

a. Use the data file *vacation* to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

b. Plot the OLS residuals versus *INCOME* and *AGE*. Do you observe any patterns suggesting that heteroskedasticity is present?

c. Sort the data according to increasing magnitude of income. Estimate the model using the first 90 observations and again using the last 90 observations. Carry out the Goldfeld–Quandt test for heteroskedastic errors at the 5% level. State the null and alternative hypotheses.

d. Estimate the model by OLS using heteroskedasticity robust standard errors. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How does this interval estimate compare to the one in (a)?

e. Obtain GLS estimates assuming $\sigma_i^2 = \sigma^2 INCOME_i^2$. Using both conventional GLS and robust GLS standard errors, construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How do these interval estimates compare to the ones in (a) and (d)?

8.17 In this exercise, we explore the relationship between total household expenditures and expenditures on clothing. Use the data file *malawi_small* (*malawi* has more observations) and observations for which *PCLOTHES* is positive. We consider three models:

$$PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e \quad \text{(XR8.17a)}$$

$$\ln(CLOTHES) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v \quad \text{(XR8.17b)}$$

$$CLOTHES = \gamma_1 + \gamma_2 TOTEXP + u \quad \text{(XR8.17c)}$$

a. Plot *PCLOTHES* versus ln(*TOTEXP*) and include the least squares fitted line. Calculate the point elasticity of clothing expenditures with respect to total expenditures at the means. See Exercise 4.12 for the elasticity in this model.

b. Calculate $CLOTHES = PCLOTHES \times TOTEXP$. Then plot ln(*CLOTHES*) versus ln(*TOTEXP*) and include the least squares fitted line. Calculate a 95% interval estimate of the elasticity of clothing expenditures with respect to total expenditures. Is the elasticity computed in part (a) within this interval?

c. Plot *CLOTHES* versus *TOTEXP* and include the least squares fitted line. Calculate a 95% interval estimate of the elasticity of clothing expenditures with respect to total expenditures at the means. Is the elasticity computed in part (a) within this interval?

d. Test for the presence of heteroskedasticity in each model in parts (a)–(c). Use the 1% level of significance. What are your conclusions? For which specification does heteroskedasticity seem less of an issue?

e. For the models in which heteroskedasticity was significant at the 1% level, use OLS with robust standard errors. Calculate a 95% interval estimate for the elasticity of clothing expenditures with respect to total expenditures at the means. How do the intervals compare to the ones based on conventional standard errors?

8.18 Consider the wage equation,

$$\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + \beta_5 FEMALE_i + \beta_6 BLACK + \beta_7 METRO_i + \beta_8 SOUTH_i + \beta_9 MIDWEST_i + \beta_{10} WEST + e_i$$

where *WAGE* is measured in dollars per hour, education and experience are in years, and *METRO* = 1 if the person lives in a metropolitan area. Use the data file *cps5* for the exercise.

a. We are curious whether holding education, experience, and *METRO* equal, there is the same amount of random variation in wages for males and females. Suppose $\text{var}(e_i|\mathbf{x}_i, FEMALE = 0) = \sigma_M^2$ and $\text{var}(e_i|\mathbf{x}_i, FEMALE = 1) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Carry out a Goldfeld–Quandt test of the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. Estimate the model by OLS. Carry out the NR^2 test using the right-hand-side variables *METRO*, *FEMALE*, *BLACK* as candidates related to the heteroskedasticity. What do we conclude about heteroskedasticity, at the 1% level? Do these results support your conclusions in (a)? Repeat the test using all model explanatory variables as candidates related to the heteroskedasticity.

c. Carry out the White test for heteroskedasticity. What is the 5% critical value for the test? What do you conclude?

d. Estimate the model by OLS with White heteroskedasticity robust standard errors. Compared to OLS with conventional standard errors, for which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

e. Obtain FGLS estimates using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to OLS with robust standard errors, from part (d)?

f. Obtain FGLS estimates with robust standard errors using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to those in part (e) and OLS with robust standard errors, from part (d)?

g. If reporting the results of this model in a research paper which one set of estimates would you present? Explain your choice.

8.19 In this exercise we explore the relationship between total household expenditures and expenditures on telephone services. Use the data file *malawi_small* (*malawi* has more observations).

a. Using observations for which $PTELEPHONE > 0$, create the variable $\ln(TELEPHONE) = \ln(PTELEPHONE \times TOTEXP)$. Plot $\ln(TELEPHONE)$ versus $\ln(TOTEXP)$ and include the least squares fitted line.

b. Based on the OLS regression of $\ln(TELEPHONE)$ on $\ln(TOTEXP)$ what is the estimated elasticity of telephone expenditures with respect to total expenditure. Compute a 95% interval estimate for the elasticity. Based on the estimates, would you classify telephone services as a necessity or a luxury?

c. Test for the presence of heteroskedasticity in the regression in part (b). What do you conclude?

d. Estimate the model $PTELEPHONE_i = \beta_1 + \beta_2 \ln(TOTEXP_i) + e_i$ by OLS. Test the null hypothesis that $\beta_2 \leq 0$ against $\beta_2 > 0$ using the 5% level of significance.

e. Calculate the elasticity of telephone expenditures with respect to total expenditure at the sample median of total expenditures. The expression for an elasticity in such a model was derived in Exercise 4.12. Use your software to compute a 95% interval estimate for the elasticity. Compare the estimated elasticity to that in (b).

f. Test for the presence of heteroskedasticity in the regression in part (d). What do you conclude?

g. Estimate the model in (d) using FGLS with $\ln(TOTEXP_i)$ being the variable that may be associated with the heteroskedasticity. Using the conventional FGLS standard errors, test the null hypothesis that $\beta_2 \leq 0$ against $\beta_2 > 0$ using the 5% level of significance.

h. Repeat part (g) but using FGLS with robust standard errors.

i. Summarize your findings about the elasticity of telephone services expenditure with respect to total expenditure.

8.20 The data file *br2* contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. We will be concerned with the selling price (*PRICE*), the size of the house in square feet (*SQFT*), the age of the house in years (*AGE*), whether the house is on a waterfront (*WATERFRONT* = 1, 0), and if it is of a traditional style (*TRADITIONAL* = 1, 0).

a. Find OLS estimates of the following equation and save the residuals.

$$\ln(PRICE) = \beta_1 + \beta_2 \ln(SQFT) + \beta_3 AGE + \beta_4 AGE^2 + \beta_5 WATERFRONT + \beta_6 TRADITIONAL + e$$

At some point, is it possible that an old house will become "historic" with age increasing its value? Construct a 95% interval estimate for the age at which age begins to have a positive effect on price.

b. Use the NR^2 test for heteroskedasticity with the candidate variables AGE, AGE^2, $WATERFRONT$, and $TRADITIONAL$. Repeat the test dropping AGE, but keeping AGE^2. Plot the least residuals against AGE. Is there any visual evidence of heteroskedasticity?

c. Estimate the model in (a) by OLS with White heteroskedasticity robust standard errors. Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the one in (a)?

d. Assume $\sigma_i^2 = \sigma^2 \exp\left(\alpha_2 AGE_i^2 + \alpha_3 WATERFRONT_i + \alpha_4 TRADITIONAL\right)$. Obtain FGLS estimates of the model in (a) and compare the results to those in (c). Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the one in (c)?

e. Obtain the residuals from the transformed model based on the skedastic function in (d). Regress the squares of these residuals on AGE^2, $WATERFRONT$, $TRADITIONAL$, and a constant term. Using the NR^2, is there any evidence of remaining heteroskedasticity in the transformed model? Repeat the test using the transformed model version of the variables and a constant term. How do the results compare?

f. Modify the estimation in (d) to use FGLS with heteroskedasticity robust standard errors. Construct a 95% interval estimate for the age at which age begins to have a positive effect on price. How does the interval compare to the ones in (c) and (d)?

g. What do you conclude about the age at which historical value increases a house price?

8.21 In Example 8.9 we estimated the linear probability model

$$COKE = \beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI + e$$

where $COKE = 1$ if a shopper purchased Coke and $COKE = 0$ if a shopper purchased Pepsi. The variable $PRATIO$ was the relative price ratio of Coke to Pepsi and $DISP\_COKE$ and $DISP\_PEPSI$ were indicator variables equal to one if the relevant display was present. Suppose now that we have 1140 observations on randomly selected shoppers from 50 different grocery stores. Each grocery store has its own settings for $PRATIO$, $DISP\_COKE$ and $DISP\_PEPSI$. Let an (i, j) subscript denote the jth shopper at the ith store, so that we can write the model as

$$COKE_{ij} = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + e_{ij}$$

Average this equation over all shoppers in the ith store so that we have

$$\overline{COKE}_i\cdot = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + \bar{e}_i\cdot \qquad \text{(XR8.21)}$$

where

$$\bar{e}_i\cdot = \frac{1}{N_i}\sum_{j=1}^{N_i} e_{ij} \qquad \text{and} \qquad \overline{COKE}_i\cdot = \frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij}$$

and N_i is the number of sampled shoppers in the ith store.

a. What is the interpretation of $\overline{COKE}_i\cdot$ for the ith store?

b. Assume that $E\left(COKE_{ij}|\mathbf{x}_{ij}\right) = P_i$ and $\text{var}\left(COKE_{ij}|\mathbf{x}_{ij}\right) = P_i\left(1 - P_i\right)$, show that $E\left(\overline{COKE}_i\cdot|\mathbf{X}\right) = P_i$ and $\text{var}\left(\overline{COKE}_i\cdot|\mathbf{X}\right) = P_i\left(1 - P_i\right)/N_i$.

c. Interpret P_i and express it in terms of $PRATIO_i$, $DISP\_COKE_i$, and $DISP\_PEPSI_i$.

d. Observations on the variables $\overline{COKE}_i\cdot$, $PRATIO_i$, $DISP\_COKE_i$, $DISP\_PEPSI_i$, and N_i appear in the data file *coke_grouped*. Obtain summary statistics for the data. Calculate the sample coefficient of variation, $\text{CV} = 100 s_x/\bar{x}$, for $\overline{COKE}_i\cdot$ and $PRATIO_i$. How much variation is there in these variables relative to their mean? Would we prefer larger or smaller coefficients of variation in these variables? Why? Construct histograms for $\overline{COKE}_i\cdot$ and $PRATIO_i$. What do you observe?

e. Find least squares estimates of equation (XR8.21) and use robust standard errors. Summarize the results. Test the null hypothesis $\beta_3 = -\beta_4$. Choose an appropriate alternative hypothesis and use the 5% level of significance. If the null hypothesis is true, what does it imply about the effect of store displays for *COKE* and *PEPSI*?

f. Create the variable $DISP = DISP\_COKE - DISP\_PEPSI$. Estimate the model $\overline{COKE}_i. = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP_i + \bar{e}_i.$ by OLS. Test for heteroskedasticity by applying the White test. Also carry out the NR^2 test for heteroskedasticity using the candidate variable N_i. What are your conclusions, at the 5% level?

g. Obtain the fitted values from (e), p_i, and estimate $\text{var}\left(\overline{COKE}_i.\right)$ for each of the stores. Report the mean, standard deviation, maximum and minimum values of the p_i.

h. Find generalized least squares estimates of the model in part (f). Comment on the results and compare them with those obtained in part (f). How might the results of part (d) help you?

8.22 Use data file *cps5* for this exercise.

a. Estimate the following wage equation by OLS and use heteroskedasticity robust standard errors:

$$\begin{aligned}\ln(WAGE) = {} & \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER \times EDUC) \\ & + \beta_6 FEMALE + \beta_7 BLACK + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e\end{aligned} \tag{XR8.22}$$

Discuss the results.

b. Add *MARRIED* to the equation and reestimate. Holding education and experience constant, do white male married workers in the northeast get higher wages? Using a 5% significance level, test a null hypothesis that wages of married workers are less than or equal to those of unmarried workers against the alternative that wages of married workers are higher.

c. Examine the residuals from part (a) for the two values of *MARRIED*. Is there evidence of heteroskedasticity?

d. Estimate the model in part (a) twice—once using observations on only married workers and once using observations on only unmarried workers. Use the Goldfeld–Quandt test and a 5% significance level to test whether the error variances for married and unmarried workers are different.

e. Hypothesize that $\sigma_i^2 = \sigma^2 \exp(\alpha_2 MARRIED)$. Find generalized least squares of the model in part (a). Compare the estimates and standard errors with those obtained in part (a).

f. Find two 95% interval estimates for the marginal effect $\partial E(\ln(WAGE)) / \partial EDUC$ for a white male worker living in the northeast with 16 years of education and 10 years of experience. Use the results from part (a) for one interval and the results from part (e) for the other interval. Comment on any differences.

8.23 Using the data in *cps5* obtain OLS estimates of the wage equation

$$\begin{aligned}\ln(WAGE) = {} & \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER \times EDUC) \\ & + \beta_6 FEMALE + \beta_7 BLACK + \beta_8 UNION + \beta_9 METRO \\ & + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e\end{aligned} \tag{XR8.23}$$

a. Interpret the coefficient of *UNION*. Test the null hypothesis that the coefficient of *UNION* is less than or equal to zero, against the alternative that is positive. What do you conclude?

b. Test for the presence of heteroskedasticity related to the variables *UNION* and *METRO* using the NR^2 test. What do you conclude at the 1% level of significance?

c. Regress the squared least squares residuals, $\hat{e}_i^2$, from (a) on *EDUC*, *UNION*, and *METRO*. Also regress $\ln\left(\hat{e}_i^2\right)$ on *EDUC*, *UNION*, and *METRO*. What do these results suggest about the effect of *UNION* membership on the variation in the random error? What do these results suggest about the effect of *METRO* on the variation in the random error?

d. Hypothesize that $\sigma_i^2 = \sigma^2 \exp(\alpha_2 EDUC + \alpha_3 UNION + \alpha_4 METRO)$. Find generalized least squares estimates of the wage equation. For the coefficient of *UNION*, compare the estimates and standard errors with those obtained from OLS estimation of (XR8.23) with heteroskedasticity robust standard errors.

8.24 In this exercise, we will explore some of the factors predicting costs at American universities using the data file *poolcoll2*. Let *TC* = the real (2008 dollars) total cost per student, *FTUG* = number of full-time undergraduate students, *FTGRAD* = number of full-time graduate students, *FTEF* = full-time faculty per 100 students, *CF* = number of contract faculty per 100 students, *FTENAP* = full

time nonacademic professionals per 100 students, *PRIVATE* = 1 if the school is private, and 0 if it is public.

a. Estimate the regression of ln(*TC*) on the remaining variables. What are the predicted effects of additional graduate students on total cost per student? What are the predicted effects of additional full-time faculty?

b. Include in the model not only *PRIVATE* but also *PRIVATE* × *FTEF*. Are these variables individually and jointly significant at the 5% level?

c. Use the NR^2 test for heteroskedasticity that is possibly related to *PRIVATE*. What do you conclude at the 1% level of significance?

d. Test the hypothesis in (b) using OLS estimates with robust standard errors.

e. Include in the model not only *PRIVATE* but also *PRIVATE* times all the other variables. Test the joint significance of *PRIVATE* and *PRIVATE* times all the other variables using an *F*-test. Use robust standard errors and carry out a robust *F*-test. Can we say "We reject the hypothesis that the models determining total cost per student are the same for public and private universities?"

f. Hypothesize $\sigma_i^2 = \exp(\alpha_1 + \alpha_2 PRIVATE)$. Obtain FGLS estimates of the model in (e) and carry out the *F*-test on *PRIVATE* and *PRIVATE* times all the other variables. What is the value of the *F*-test statistic? What is the 1% critical value?

8.25 What effect does having public health insurance have on the number of doctor visits a person has during a year? Using 1988 data, in the data file *rwm88_small*, from Germany, we will explore this question. The data file *rwm88* contains more observations.

a. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *PUBLIC*, *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*. Test the null hypothesis that the coefficient on *PUBLIC* is less than or equal to zero, versus the alternative that it is greater than zero at the 1% level of significance.

b. Test for the presence of heteroskedasticity. Obtain the squared least squares residuals from the regression in (a), regress them on all the explanatory variables, and carry out an *F*-test of their joint significance. What do we conclude about the presence of heteroskedasticity at the 1% level of significance?

c. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* separately for those with public insurance and those who do not have public insurance. Use equation (7.37) to obtain the estimate of the average treatment effect of public insurance.

d. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *PUBLIC*, *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2* in "deviation from the mean" form. That is, for each variable x, create the variable $\tilde{x} = x - \bar{x}$, where $\bar{x}$ is the sample mean. Using robust standard errors, test the significance of the coefficient on *PUBLIC*.

e. Estimate the regression model with the dependent variable *DOCVIS* and the explanatory variables *FEMALE*, *HHKIDS*, *MARRIED*, *SELF*, *EDUC2*, *HHNINC2*, along with *PUBLIC* and *PUBLIC* times each of the variables in deviation about the mean form. What is the estimated average treatment effect? Using a robust standard error, is it statistically significant at the 5% level? [*Hint*: See equation (7.41) and the surrounding discussion.]

8.26 In the STAR experiment, Example 7.8, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular–sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes is contained in the data file *star5_small2*.

a. Regress *MATHSCORE* on *SMALL*, *AIDE*, *TCHEXPER*, *SCHRURAL*, *FREELUNCH*, and *BOY*. Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

b. Estimate the regression model in (a) by OLS including interactions between *FREELUNCH* and the other variables. Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

c. Using the model in (b), and both conventional and robust standard errors, test the joint significance of the interactions between *FREELUNCH* and *SMALL*, *AIDE*, and *TCHEXPER* at the 10% level in each regression. What do you conclude?

d. Estimate the model in (b) and include indicator variables for each school (*SCHOOLID*). Test for heteroskedasticity related to *SMALL* and *AIDE* using the NR^2 test. What do you conclude at the 5% level?

e. Using the model in (d), and both conventional and robust standard errors, test the joint significance of the interactions between *FREELUNCH* and *SMALL*, *AIDE*, and *TCHEXPER* at the 10% level in each regression. What do you conclude?

8.27 There were 64 countries who competed in the 1992 Olympics and won at least one medal. For each of these countries, let *MEDALTOT* be the total number of medals won, *POP* be population in millions, and *GDP* be GDP in billions of 1995 dollars.

a. Use the data file *olympics5*, excluding the United Kingdom, and use the $N = 63$ remaining observations. Estimate the model $MEDALTOT = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + e$ by OLS.

b. Calculate the squared least squares residuals $\hat{e}_i^2$ from the regression in (a). Regress $\hat{e}_i^2$ on $\ln(POP)$ and $\ln(GDP)$. Use the F-test from this regression to test for heteroskedasticity at the 5% level of significance. Use the R^2 from this regression to test for heteroskedasticity. What are the p-values of the two tests?

c. Reestimate the model in (a) but using heteroskedasticity robust standard errors. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

d. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

e. Use the model in (c) to find point and 95% interval estimates for the expected number of medals won by the United Kingdom whose population and GDP in 1992 were 58 million and $1010 billion, respectively.

f. The United Kingdom won 20 medals in 1992. Was the model successful in predicting the mean number of medals for the United Kingdom? Using the estimation in (c), with robust standard errors, what is the p-value for a test of $H_0 : \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 = 20$ versus $H_1 : \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 \neq 20$?

8.28 In this exercise you will create some simulated data and try out estimation and testing methods. Use your software to create a new data set, or "workfile," with $N = 100$ observations. All modern software has functions, called random number generators, to create uniformly distributed and normally distributed random values. Follow these steps.

1. Create $X2 = 1 + 5 \times U1$, where $U1$ is a random number between zero and one.
2. Create $X3 = 1 + 5 \times U2$, where $U2$ is another random number between zero and one.
3. Create $E = \sqrt{\exp(2 + 0.6X2)} \times Z$, where $Z \sim N(0, 1)$.
4. Create $Y = 5 + 4X2 + E$

You should now have 100 values for Y, $X2$, and $X3$. Note: Your results should be different from your classmates, and your results might change from one experiment to the next. To prevent this from happening, you can set the random number's "seed." See your software documentation for instructions.

a. Regress Y on $X2$ and $X3$ and obtain conventional OLS standard errors. Compare the estimated coefficients to the true values of the regression parameters, $\beta_1 = 5$, $\beta_2 = 4$, $\beta_3 = 0$. Do the t-values suggest that the coefficients are significantly different from 0 at the 5% level?

b. Calculate the least squares residuals $\hat{e}$ from the OLS estimation in (a) and regress $\hat{e}^2$ on $X2$ and $X3$. What evidence, if any, do you find for the presence of heteroskedasticity?

c. Regress Y on $X2$ and $X3$ and obtain robust standard errors. Compare these to the conventional standard errors in (a).

d. Assume the heteroskedasticity pattern is $\sigma^2 X2^2$. Obtain GLS estimates with conventional and robust standard errors. Are the GLS parameter estimates closer to the true parameter values or not? Which set of standard errors should be used?

e. Assume the multiplicative heteroskedasticity model $\exp(\alpha_1 + \alpha_2 X2 + \alpha_3 X3)$. Obtain FGLS estimates with conventional and robust standard errors. Are the FGLS estimates closer to the true parameter values than the GLS or OLS estimates? Which set of standard errors should be used?

8.29 The data file *mexican* contains data collected in 2001 from the transactions of 754 Mexican sex workers.

- **a.** Using OLS, estimate the hedonic log-linear model with *LNPRICE* as the dependent variable and independent variables *BAR*, *STREET*, *SCHOOL*, *AGE*, *RICH*, *ALCOHOL*, *ATTRACTIVE*. Interpret the estimated coefficients.
- **b.** Test for heteroskedasticity related to *ATTRACTIVE* using the NR^2 test at the 1% level of significance.
- **c.** Estimate the model separately by OLS for observations with *ATTRACTIVE* = 1 and *ATTRACTIVE* = 0. Using the results, carry out the Goldfeld–Quandt test for heteroskedasticity across the two regressions. Use a two-tailed test at the 5% level. Which regression has a larger estimated error variance?
- **d.** Compare the estimates from the two estimations in (c). Do they appear similar or dissimilar? Which coefficients are noticeably different? Use OLS to estimate the model that includes the original variables and interactions between *ATTRACTIVE* and the other explanatory variables. Test the joint significance of *ATTRACTIVE* and the interaction variables at the 1% level of significance. Is this a "valid" Chow test? Is homoskedasticity a necessary condition for this test? Recall that the test is described in Section 7.2.3.
- **e.** Using the estimation results in (d), test for heteroskedasticity related to *ATTRACTIVE* using the NR^2 test at the 1% level of significance.
- **f.** Use OLS with heteroskedasticity robust standard errors to estimate the model that includes the original variables and interactions between *ATTRACTIVE* and the other explanatory variables. Test the joint significance of *ATTRACTIVE* and the interaction variables at the 1% level of significance. Is this a "valid" Chow test?

8.30 The data file *grunfeld2* contains annual data on the gross investment, capital stock, and the value of the firm, measured by the value of common and preferred stock for General Electric and Westinghouse, during the period 1935–1954. These data have been used to train econometricians for almost 60 years, and still provide valuable lessons.

- **a.** Create an indicator variable $GE = 1$ for General Electric and $GE = 0$ for Westinghouse. Using the combined data on both firms, use OLS to estimate the model of investment, *INV*, as a function of the value of the firms, *V*, and capital stock, *K*, also the indicator variable *GE* and the interactions of *GE* with *V* and *K*. That is $INV = f(const, V, K, GE, GE \times V, GE \times K)$. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. What does this test reveal about the two firms' investment characteristics?
- **b.** Obtain the OLS residuals from (a) and regress their squares on the indicator variable *GE*. Use the result of this regression to test for heteroskedasticity across the firms at the 1% level.
- **c.** Reestimate the model in (a) using OLS with heteroskedasticity robust standard errors. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. Does your conclusion change?
- **d.** Estimate the investment model separately for General Electric and Westinghouse. Let the estimated error variances be $\hat{\sigma}^2_{GE}$ and $\hat{\sigma}^2_{WE}$. For which firm is the estimated error variance smaller?
- **e.** Create a variable *W* that takes the value $\hat{\sigma}^2_{GE}$ when $GE = 1$ and takes the value $\hat{\sigma}^2_{WE}$ when $GE = 0$. Estimate the model in (a) by FGLS with weighting variable *W*. Test the joint significance of the variables $GE, GE \times V, GE \times K$ at the 5% level. Does your conclusion change?

Appendix 8A Properties of the Least Squares Estimator

In Appendix 2D, we wrote the least squares estimator for β_2 in the simple regression model as $b_2 = \beta_2 + \sum w_i e_i$, where

$$w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

This expression is a useful one for exploring the properties of the least squares estimator under heteroskedasticity. The first property that we establish is that of unbiasedness. This property was

derived under homoskedasticity in equation (2.13) of Chapter 2. The same proof holds under heteroskedasticity because the only error term assumption that was used is $E(e_i|\mathbf{x}) = 0$.

$$\begin{aligned} E(b_2|\mathbf{x}) &= E(\beta_2 + \sum w_i e_i|\mathbf{x}) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N|\mathbf{x}) \\ &= E(\beta_2) + E(w_1 e_1|\mathbf{x}) + E(w_2 e_2|\mathbf{x}) + \cdots + E(w_N e_N|\mathbf{x}) \\ &= \beta_2 + \sum E(w_i e_i|\mathbf{x}) = \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 \end{aligned}$$

The least squares estimators are unbiased as long as $E(e_i|\mathbf{x}) = 0$, even if the errors are heteroskedastic. This is true in both the simple and multiple regression models.

The variance of the least squares estimator is

$$\begin{aligned} \text{var}(b_2|\mathbf{x}) &= \text{var}(\sum w_i e_i|\mathbf{x}) \\ &= \sum w_i^2 \text{var}(e_i|\mathbf{x}) + \sum_{i \neq j} \sum w_i w_j \text{cov}(e_i, e_j|\mathbf{x}) \\ &= \sum w_i^2 \sigma_i^2 \\ &= \sum \left\{ \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right\}^2 \sigma_i^2 = \sum \left\{ \frac{(x_i - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} \sigma_i^2 \right\} \\ &= \left[\sum (x_i - \bar{x})^2\right]^{-1} \sum \left[(x_i - \bar{x})^2 \sigma_i^2\right] \left[\sum (x_i - \bar{x})^2\right]^{-1} \end{aligned} \tag{8A.1}$$

Going from the second line to the third we used assumption MR4, conditionally uncorrelated errors, $\text{cov}(e_i, e_j|\mathbf{x}) = 0$. If the variances are all the same $(\sigma_i^2 = \sigma^2)$, then the third line becomes $\sigma^2 \sum w_i^2 = \text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum (x_i - \bar{x})^2$, which is the usual OLS variance expression. This simplification is not possible under heteroskedasticity. The fourth and fifth lines are equivalent ways of writing the variance of the least squares estimator, equation (8.8), when the random errors are heteroskedastic.

Appendix 8B Lagrange Multiplier Tests for Heteroskedasticity

More insights into LM and other variance function tests can be developed by relating them to the F-test introduced in (6.8) for testing the significance of a mean function. To put that test in the context of a variance function, consider (8.15)

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS} + v_i \tag{8B.1}$$

and assume that our objective is to test $H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$ against the alternative that at least one α_s, for $s = 2, \ldots, S$, is nonzero. In Section 8.2.2 we considered a more general variance function than that in (8B.1), but we also pointed out that using the linear function in (8B.1) is valid for testing more general alternative hypotheses.

Adapting the F-value reported in (6.8) to test the overall significance of (8B.1), we have

$$F = \frac{(SST - SSE)/(S-1)}{SSE/(N-S)} \tag{8B.2}$$

where

$$SST = \sum_{i=1}^{N} \left[\hat{e}_i^2 - \overline{\hat{e}^2}\right]^2 \quad \text{and} \quad SSE = \sum_{i=1}^{N} \hat{v}_i^2$$

are the total sum of squares and sum of squared errors from estimating (8B.1). Note that $\overline{\hat{e}^2}$ is the mean of the dependent variable in (8B.1), or, equivalently, the average of the squares of the least squares residuals from the regression function. At a 5% significance level, a valid test is to reject H_0 if the F-value is greater than a critical value given by $F_{(0.95,\, S-1,\, N-S)}$.

Two further tests, the original Breusch–Pagan test and its $N \times R^2$ version, can be obtained by modifying (8B.2). Please be patient as we work through these modifications. We begin by rewriting (8B.2) as

$$\chi^2 = (S-1) \times F = \frac{SST - SSE}{SSE/(N-S)} \sim \chi^2_{(S-1)} \tag{8B.3}$$

The chi-square statistic $\chi^2 = (S-1) \times F$ has an approximate $\chi^2_{(S-1)}$ -distribution in large samples. That is, multiplying an F-statistic by its numerator degrees of freedom gives another statistic that follows a chi-square distribution. The degrees of freedom of the chi-square distribution are $S-1$, the same as that for the numerator of the F-distribution. The background for this result is given in Appendix 6A.

Next, note that

$$\widehat{\text{var}}\left(e_i^2\right) = \widehat{\text{var}}\left(v_i\right) = \frac{SSE}{N-S} \tag{8B.4}$$

That is, the variance of the dependent variable is the same as the variance of the error, which can be estimated from the sum of squared errors in (8B.1). Substituting (8B.4) into (8B.3) yields

$$\chi^2 = \frac{SST - SSE}{\widehat{\text{var}}\left(e_i^2\right)} \tag{8B.5}$$

This test statistic represents the basic form of the Breusch–Pagan statistic. Its two different versions occur because of the alternative estimators used to replace $\widehat{\text{var}}\left(e_i^2\right)$.

If it is assumed that e_i is normally distributed, it can be shown that $\text{var}\left(e_i^2\right) = 2\sigma_e^4$, and the statistic for the first version of the Breusch–Pagan test is

$$\chi^2 = \frac{SST - SSE}{2\hat{\sigma}_e^4} \tag{8B.6}$$

Note that $\sigma_e^4 = \left(\sigma_e^2\right)^2$ is the square of the error variance from the mean function; unlike *SST* and *SSE*, its estimate comes from estimating (8.16). The result $\text{var}\left(e_i^2\right) = 2\sigma_e^4$ might be unexpected—here is a little proof so that you know where it comes from. When $e_i \sim N\left(0, \sigma_e^2\right)$, then $\left(e_i/\sigma_e\right) \sim N(0,1)$, and $\left(e_i^2/\sigma_e^2\right) \sim \chi^2_{(1)}$. The variance of a $\chi^2_{(1)}$ random variable is 2. Thus,

$$\text{var}\left(\frac{e_i^2}{\sigma_e^2}\right) = 2 \Rightarrow \quad \frac{1}{\sigma_e^4}\text{var}\left(e_i^2\right) = 2 \Rightarrow \text{var}\left(e_i^2\right) = 2\sigma_e^4$$

Using (8B.6), we reject a null hypothesis of homoskedasticity when the χ^2-value is greater than a critical value from the $\chi^2_{(S-1)}$ distribution.

For the second version of (8B.5) the assumption of normally distributed errors is not necessary. Because this assumption is not used, it is often called the robust version of the Breusch–Pagan test. The sample variance of the squared least squares residuals, the $\hat{e}_i^2$, is used as an estimator for $\text{var}\left(e_i^2\right)$. Specifically, we set

$$\widehat{\text{var}}\left(e_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{e}_i^2 - \overline{\hat{e}^2}\right]^2 = \frac{SST}{N} \tag{8B.7}$$

This quantity is an estimator for $\text{var}\left(e_i^2\right)$ under the assumption that H_0 is true. It can also be written as the total sum of squares from estimating the variance function divided by the sample size.

Substituting (8B.7) into (8B.5) yields

$$\begin{aligned}\chi^2 &= \frac{SST - SSE}{SST/N} \\ &= N \times \left(1 - \frac{SSE}{SST}\right) \\ &= N \times R^2\end{aligned} \tag{8B.8}$$

where R^2 is the R^2 goodness-of-fit statistic from estimating the variance function. At a 5% significance level, a null hypothesis of homoskedasticity is rejected when $\chi^2 = N \times R^2$ exceeds the critical value $\chi^2_{(0.95,S-1)}$.

Software often reports the outcome of the White test described in Section 8.6.4 a as an F-value or a χ^2-value. The F-value is from the statistic in (8B.4), with the z's chosen as the x's and their squares and possibly cross-products. The χ^2-value is from the statistic in (8B.8), with the z's chosen as the x's and their squares and possibly cross-products.

Appendix 8C Properties of the Least Squares Residuals

The least squares residuals are $\hat{e}_i = y_i - \hat{y}_i$. Substituting in the fitted value $\hat{y}_i = b_1 + b_2 x_i$ we obtain for the simple regression model

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + e_i - (b_1 + b_2 x_i) \\ &= (\beta_1 - b_1) + (\beta_2 - b_2) x_i + e_i \\ &= e_i - (b_1 - \beta_1) - (b_2 - \beta_2) x_i\end{aligned}$$

Using the last line we find

$$E(\hat{e}_i|\mathbf{x}) = E(e_i|\mathbf{x}) - E(b_1 - \beta_1|\mathbf{x}) - E(b_2 - \beta_2|\mathbf{x}) x_i = 0$$

The expected value of the least squares residual is zero under assumptions SR1–SR5. Also, note what happens if we consider large samples, with $N \to \infty$. The least squares estimators b_1 and b_2 are unbiased, and recall from Section 2.4.4 that their variances get smaller and smaller as N gets larger. This means that in large samples $(b_1 - \beta_1)$ and $(b_2 - \beta_2)$ are close to zero, so that in large samples the difference $\hat{e}_i - e_i$ is close to zero. In econometric terms, the *probability limit* of $\hat{e}_i - e_i$ is zero, that is, $\text{plim}(\hat{e}_i - e_i) = 0$. The two random variables become essentially the same and thus have the same probability distribution. This means, that in large samples, if $e_i \sim N(0, \sigma^2)$ then $\hat{e}_i \overset{a}{\sim} N(0, \sigma^2)$, where "$\overset{a}{\sim}$" means **approximately distributed**, or **asymptotically** (in large samples) **distributed**. Learning asymptotic analysis is an important feature of econometrics. See Section 5.7 for further discussion.

It can be shown that the conditional variance of the least squares residual is

$$\text{var}(\hat{e}_i|\mathbf{x}) = E\left(\hat{e}_i^2|\mathbf{x}\right) = \sigma^2 \left\{1 - \frac{1}{N} - \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right\} = \sigma^2 (1 - h_i) \tag{8C.1}$$

where h_i is the **leverage** of the ith observation, a term we introduced in Section 4.3.6. Note that:

- **i.** The conditional variance of the least squares residual is not constant even if the random error is homoskedastic.
- **ii.** Because $0 \le h_i \le 1$ and $0 \le (1 - h_i) \le 1$, $\text{var}(\hat{e}_i|\mathbf{x}) < \text{var}(e_i|\mathbf{x}) = \sigma^2$. The variation in the least squares residual is less than the variance of the true random error.
- **iii.** The variance of the least squares residual is closest to $\text{var}(e_i|\mathbf{x}) = \sigma^2$ when $x_i = \bar{x}$, reflecting the fact that the fitted value $\hat{y}_i$ has the least prediction error at that point.
- **iv.** The expression (8C.1) is valid in both simple and multiple regression, with h_i redefined in multiple regression.

v. The sum of the leverage values is K, $\sum h_i = K$. As a check, verify that for the simple regression model $\sum h_i = 2$.

vi. $\sum_{i=1}^{N} \text{var}(\hat{e}_i|\mathbf{x}) = \sum_{i=1}^{N} E\left(\hat{e}_i^2|\mathbf{x}\right) = \sigma^2(N-K)$ while $\sum_{i=1}^{N} \text{var}(e_i|\mathbf{x}) = \sum_{i=1}^{N} E(e_i^2|\mathbf{x}) = N\sigma^2$

8C.1 Details of Multiplicative Heteroskedasticity Model

We showed that the least squares residuals and the true random error have the same probability distribution in large samples. If $e_i \sim N(0, \sigma_i^2)$ then in large samples the least squares residual $\hat{e}_i \overset{a}{\sim} N(0, \sigma_i^2)$. In large samples, then $(\hat{e}_i/\sigma_i) \overset{a}{\sim} N(0,1)$ and $(\hat{e}_i/\sigma_i)^2 \overset{a}{\sim} [N(0,1)]^2 \sim \chi^2_{(1)}$. Thus,

$$\ln\left[(\hat{e}_i/\sigma_i)^2\right] = v_i \overset{a}{\sim} \ln\left[\chi^2_{(1)}\right]$$

Statisticians have studied this random variable and found that $E\left\{\ln\left[\chi^2_{(1)}\right]\right\} = -1.2704$ and $\text{var}\left\{\ln\left[\chi^2_{(1)}\right]\right\} = 4.9348$.

Appendix 8D Alternative Robust Sandwich Estimators

The robust variance estimators carry over to the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ quite easily. Recall from Appendix 6B that we can express the least squares estimator b_2 as

$$b_2 = \frac{\sum(x_{i2} - \tilde{x}_{i2})\, y_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}$$

where $\tilde{x}_{i2}$ is the fitted value from the auxiliary regression of x_2 on all the other explanatory variables, $x_{i2} = c_1 + c_3 x_{i3} + \cdots + c_K x_{iK} + r_{i2}$. Substituting for y_i and simplifying leads us to

$$b_2 = \beta_2 + \frac{\sum(x_{i2} - \tilde{x}_{i2})\, e_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}$$

If the errors are heteroskedastic and serially uncorrelated, then the conditional variance of b_2 is

$$\begin{aligned} \text{var}(b_2|\mathbf{X}) &= \text{var}\left[\left.\frac{\sum(x_{i2} - \tilde{x}_{i2})\, e_i}{\sum(x_{i2} - \tilde{x}_{i2})^2}\right| \mathbf{X}\right] = \frac{\sum(x_{i2} - \tilde{x}_{i2})^2 \text{var}(e_i|\mathbf{X})}{\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^2} \\ &= \frac{\sum(x_{i2} - \tilde{x}_{i2})^2 \sigma_i^2}{\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^2} \\ &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \left\{\sum(x_{i2} - \tilde{x}_{i2})^2 \sigma_i^2\right\} \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \end{aligned} \tag{8D.1}$$

The **original** White heteroskedasticity corrected variance estimator replaces σ_i^2 by the squared OLS residuals

$$\widehat{\text{var}}(b_2) = \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \left\{\sum(x_{i2} - \tilde{x}_{i2})^2 \hat{e}_i^2\right\} \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} = \text{HCE0} \tag{8D.2}$$

The version in equation (8D.2) is valid in large samples. In practice, some alternatives are used that are designed to work better in smaller samples. These alternatives account for the fact that the least squares residuals are on average a little smaller than the true random errors. As noted in

Appendix 8C, in the simple regression model, with assumptions SR1–SR5 holding, the variance of the least squares residual is

$$\text{var}(\hat{e}_i|\mathbf{X}) = E\left(\hat{e}_i^2|\mathbf{X}\right) = \sigma^2\left\{1 - \frac{1}{N} - \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right\} = \sigma^2(1 - h_i) \tag{8D.3}$$

where h_i is the **leverage** of the ith observation, a term we introduced in Section 4.3.6. In the simple regression model

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

The expression

$$\text{var}(\hat{e}_i|\mathbf{X}) = \sigma^2(1 - h_i) \tag{8D.4}$$

is valid in both simple and multiple regression, with h_i redefined when $K > 2$. For both simple and multiple regression, $0 \le h_i \le 1$ and $0 \le (1 - h_i) \le 1$.

The first modification of HCE0 is based on the observation that the expected value of the squared least squares residual is smaller than the expected value of the squared random errors.

$$\text{var}(\hat{e}_i|\mathbf{X}) = E\left(\hat{e}_i^2|\mathbf{X}\right) = \sigma^2(1 - h_i) < \text{var}(e_i|\mathbf{X}) = E(e_i^2|\mathbf{X}) = \sigma^2$$

The average value of $E\left(\hat{e}_i^2|\mathbf{X}\right)$ is $[(N - K)/N]\sigma^2$ while the average value of $E(e_i^2|\mathbf{X}) = \sigma^2$. To adjust for the size difference of the least squares residuals, multiply $\hat{e}_i^2$ in HCE0 by $N/(N - K)$. That is,

$$\begin{aligned}\widehat{\text{var}}(b_2) &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2} - \tilde{x}_{i2})^2\left(\frac{N}{N - K}\right)\hat{e}_i^2\right]\right\}\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \\ &= \text{HCE1}\end{aligned} \tag{8D.5}$$

This correction will have little effect if the sample is large, but it may have an effect when the number of explanatory variables in the model, $K - 1$, is large.

A second modification adjusts the squared least squares residual to have the same conditional expectation as the random error. That is,

$$E\left(\left.\frac{\hat{e}_i^2}{1 - h_i}\right|\mathbf{X}\right) = \sigma^2 = E(e_i^2|\mathbf{X})$$

Then, HCE2 is

$$\begin{aligned}\widehat{\text{var}}(b_2) &= \left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2} - \tilde{x}_{i2})^2\frac{\hat{e}_i^2}{(1 - h_i)}\right]\right\}\left[\sum(x_{i2} - \tilde{x}_{i2})^2\right]^{-1} \\ &= \text{HCE2}\end{aligned} \tag{8D.6}$$

In large samples HCE0, HCE1, and HCE2 are equivalent, but in samples that are not very large, the adjustments make useful differences. In econometric software, the "default" robust variance estimator is HCE0 or HCE1. If the random errors are actually homoskedastic, using HCE2 seems appropriate. Recall that part of the genius of the White heteroskedasticity robust variance estimators is that in large samples they can be applied whether the random errors are heteroskedastic or not. The modification introduced in HCE2 "tweaks" the robust estimator in such a way that it works when the errors are heteroskedastic and a little better than HCE0 and HCE1 when errors are homoskedastic.

Recall that $0 \le (1 - h_i) \le 1$ so HCE2 inflates the least squares residuals and the larger the leverage, h_i, the larger the adjustment becomes. Observations with high leverage, ones that have

a larger impact on regression estimates and predictions, are also the observations for which the least squares residual is much too small, thus the third modification inflates the residual again, using

$$\frac{\hat{e}_i^2/(1-h_i)}{(1-h_i)} = \frac{\hat{e}_i^2}{(1-h_i)^2}$$

Then

$$\widehat{\text{var}}(b_2) = \left[\sum(x_{i2}-\tilde{x}_{i2})^2\right]^{-1}\left\{\sum\left[(x_{i2}-\tilde{x}_{i2})^2\frac{\hat{e}_i^2}{(1-h_i)^2}\right]\right\}\left[\sum(x_{i2}-\tilde{x}_{i2})^2\right]^{-1}$$
$$= \text{HCE3} \tag{8D.7}$$

Some research shows that if heteroskedasticity is present in the data, then HCE3 is a good choice.

To summarize, replacing σ_i^2 in (8D.1) by $\hat{e}_i^2$, $[N/(N-K)]\hat{e}_i^2$, $\hat{e}_i^2/(1-h_i)$, or $\hat{e}_i^2/(1-h_i)^2$ leads to the robust sandwich variance estimators HCE0, HCE1, HCE2, or HCE3. These robust sandwich variance estimators are equivalent in large samples but may yield different results in small samples. "Robust" means that the variance estimates, and standard errors, are valid whether heteroskedasticity is present or not. When a priori reasoning *does not* lead you to suspect heteroskedasticity, but you are suspicious and/or risk averse, and if your sample is not small, then using the robust sandwich variance estimator HCE2 may be the best choice. When a priori reasoning *does* lead you to suspect heteroskedasticity, and if your sample is not small, then using the robust sandwich variance estimator HCE3 may be the better choice. Because the calculations are complex, it is best to use proper econometric software for robust variances.

EXAMPLE 8.10 | Alternative Robust Standard Errors in the Food Expenditure Model

Most regression packages include an option for calculating standard errors using White's estimator. If we do so for the food expenditure example, we obtain

$$\widehat{FOOD\_EXP} = 83.42 + 10.21\,INCOME$$
$$(27.46)\quad (1.81)\quad \text{(White robust se-HCE1)}$$
$$(27.69)\quad (1.82)\quad \text{(White robust se-HCE2)}$$
$$(28.65)\quad (1.89)\quad \text{(White robust se-HCE3)}$$
$$(43.41)\quad (2.09)\quad \text{(incorrect OLS se)}$$

In this case, ignoring heteroskedasticity and using incorrect standard errors, based on the usual formula in (8.6), tends to understate the precision of estimation; we tend to get confidence intervals that are wider than they should be. Specifically, following the result in (3.6) in Chapter 3, we can construct four corresponding 95% confidence intervals for β_2.

White HCE1: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.81 = [6.55,\ 13.87]$

White HCE2: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.82 = [6.52,\ 13.90]$

White HCE3: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 1.89 = [6.39,\ 14.03]$

Incorrect: $b_2 \pm t_c\text{se}(b_2) = 10.21 \pm 2.024 \times 2.09 = [5.97,\ 14.45]$

If we ignore heteroskedasticity, we estimate that β_2 lies between 5.97 and 14.45. When we recognize the existence of heteroskedasticity, our information is more precise, and using HCE3 we estimate that β_2 lies between 6.39 and 14.03. Why HCE3? Because a priori we could reason that heteroskedasticity should be present. A caveat here is that the sample is small, which does mean that the robust standard error formulas we have provided may not be as accurate as if the sample were large.

Appendix 8E Monte Carlo Evidence: OLS, GLS, and FGLS

White's estimator for the standard errors helps us avoid computing incorrect interval estimates or incorrect values for test statistics in the presence of heteroskedasticity. The least squares estimator is no longer best, but failing to use the "best" estimator may not be too grave a sin if estimates are sufficiently precise for useful economic analysis. Many cross-sectional data sets have thousands of observations, resulting in robust standard errors that are small, making interval estimates narrow and *t*-tests powerful. Nothing further is required in these cases. If, however, your estimates are not sufficiently precise for economic analysis, then a better, more efficient estimator is called for. In order to use such an estimator, we must specify the **skedastic** function $h(\mathbf{x}_i) > 0$, a function of $\mathbf{x}_i$ and also perhaps other variables, that describe the pattern of conditional heteroskedasticity. In this appendix, we use a Monte Carlo study to illustrate an alternative estimator, feasible generalized least squares, that has a smaller variance than the least squares estimator in large samples.

Using Monte Carlo experiments, we illustrate the properties of the OLS estimator, the correct FGLS estimator and an incorrect GLS estimator. The data generating process[4] is based on the population model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i = 5 + x_{i2} + 0x_{i3} + e_i$$

The variables x_2 and x_3 are statistically independent uniform (Appendix B.3.4) random variables over the interval (1, 5). They vary randomly with all values being equally likely in the interval. The random error is $e_i = h(\mathbf{x}_i) z_i$, where $z_i \sim N(0, 1)$. The skedasticity function $h(\mathbf{x}_i)$ is

$$h(\mathbf{x}_i) = 3\exp(1 + \alpha_2 x_{i2} + 0x_{i3}) / \overline{h}$$

The value of α_2 changes from $\alpha_2 = 0$, homoskedasticity, to $\alpha_2 = 0.3$, strong heteroskedasticity, to $\alpha_2 = 0.5$, very strong heteroskedasticity. The scalar $\overline{h}$ is a constant such that $\sum_{i=1}^{N} h(\mathbf{x}_i)/N \cong 3$ so that $\sum_{i=1}^{N} \text{var}(e_i|\mathbf{x}_i)/N \cong 9$. We use two sample sizes, $N = 100$, a moderate sample size, and $N = 5000$, a large sample. We use $M = 1000$ Monte Carlo replications and do not hold x_2 and x_3 constant across these experiments.

In Table 8E.1 we report the results of the experiments. The FGLS procedure follows the description in Section 8.5.1, with equation (8.20) being $\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + v_i$. The GLS estimation incorrectly assumes $\text{var}(e_i|\mathbf{x}_i) = \sigma^2 x_{i2}$. This is the proportional heteroskedasticity assumption illustrated in Section 8.4.1. In the first row of Table 8E.1 is the sample size, N, and in the second row is the value of α_2. First, the results of experiments (1)–(4):

1. Let the OLS estimator of β_2 be b_2. The OLS estimator is unbiased in the presence of heteroskedasticity, which is revealed by the Monte Carlo average across 1000 samples $\overline{b}_2$ in row (3) that is close to the true value $\beta_2 = 1$. The averages of the (correct) FGLS estimates, $\overline{\hat{\beta}}_2$, in row (8) and the (incorrect) GLS estimates, $\overline{\tilde{\beta}}_2$, in row (13) are also close to the true parameter value.
2. The sample standard deviation of the 1000 Monte Carlo OLS estimates is $\text{sd}(b_2)$ in row (4). It measures the actual amount of sampling variation of the OLS estimator—how much it varies from sample to sample due solely to randomness inherent in sampling from a population. Compare to it the sample average of the 1000 Monte Carlo calculated values of the

[4]This design is adapted from James G. MacKinnon (2013) "Thirty Years of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, editors Xiaohong Chen and R. Norman Swanson, New York: Springer, 437–461.

usual, or nominal, OLS standard error of the estimator b_2, $\overline{\text{se}}(b_2)$ in row (5). Note that when $N = 100$ and $\alpha_2 \neq 0$ the average standard error is less than the standard deviation, meaning that the OLS standard error is too small on average. When $N = 5000$ both of these values are dramatically reduced, but the OLS standard error is still on average too small. Now compare the average of the White robust standard errors HCE1 with the simple inflation factor $N/(N-3)$ as described in Appendix 8C, $\overline{\text{robse}}(b_2)$ in row (6). The average of these standard errors is very close to the actual variation measured by $\text{sd}(b_2)$. That means that the robust standard error correction for the OLS estimator is doing its job, on average, in measuring actual sampling variation.

3. When heteroskedasticity is present, the actual variation in the FGLS estimates, $\text{sd}(\hat{\hat{\beta}}_2)$ in row (9) is less than the actual variation in the OLS estimates, $\text{sd}(b_2)$. The ratio $\text{sd}(\hat{\hat{\beta}}_2)/\text{sd}(b_2)$ in row (10) shows the improvement obtained by using FGLS. By using FGLS, we have obtained estimates that are more precise than the OLS estimates, as we should have. The sample average of the standard error estimates $\overline{\text{se}}(\hat{\hat{\beta}}_2)$, row (11), is slightly smaller than $\text{sd}(\hat{\hat{\beta}}_2)$ when $N = 100$. In this sample size, the FGLS standard errors are a little too small. When $N = 5000$ this is no longer the case. We are reminded that the properties of the FGLS estimator are valid in large samples. We used the correct model for the heteroskedasticity in the FGLS calculations; hence, there is no need to compute FGLS

TABLE 8E.1 **Monte Carlo Simulation Results**

		Experiment				
Result	**Item**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
1	N	100	100	100	5000	5000
2	α_2	0	0.3	0.5	0.5	NA
3	$\bar{b}_2$	1.0058	1.0044	1.0033	0.9996	1.0007
4	$\text{sd}(b_2)$	0.2657	0.3032	0.3574	0.0496	0.0414
5	$\overline{\text{se}}(b_2)$	0.2626	0.2831	0.3081	0.0423	0.0406
6	$\overline{\text{robse}}(b_2)$	0.2614	0.3035	0.3586	0.0498	0.0406
7	$\text{rej}(NR^2)$	0.0570	0.9620	1.0000	1.0000	0.0420
8	$\bar{\hat{\beta}}_2$	1.0070	1.0114	1.0116	1.0000	1.0013
9	$\text{sd}(\hat{\hat{\beta}}_2)$	0.2746	0.2731	0.2522	0.0312	0.0452
10	$\text{sd}(\hat{\hat{\beta}}_2)/\text{sd}(b_2)$	1.0338	0.9007	0.7058	0.6299	1.0920
11	$\overline{\text{se}}(\hat{\hat{\beta}}_2)$	0.2608	0.2555	0.2351	0.0323	0.0415
12	$\overline{\text{robse}}(\hat{\hat{\beta}}_2)$	0.2610	0.2565	0.2371	0.0323	0.0442
13	$\bar{\hat{\beta}}_2$	1.0124	1.0092	1.0073	0.9996	1.0007
14	$\text{sd}(\hat{\beta}_2)$	0.2924	0.2680	0.2894	0.0392	0.0414
15	$\text{sd}(\hat{\beta}_2)/\text{sd}(b_2)$	1.1009	0.8839	0.8099	0.7900	0.0406
16	$\overline{\text{se}}(\hat{\beta}_2)$	0.2677	0.2512	0.2561	0.0349	0.0406
17	$\overline{\text{robse}}(\hat{\beta}_2)$	0.2794	0.2645	0.2888	0.0395	0.0420

with robust standard errors, but we report these values for reference in row (12), $\overline{\text{robse}}\left(\hat{\hat{\beta}}_2\right)$. The averages are not much different from $\overline{\text{se}}\left(\hat{\hat{\beta}}_2\right)$, as we would have guessed.

4. The sampling variation of the GLS estimator, $\text{sd}\left(\hat{\beta}_2\right)$, is in row (14). The average of the usual, or nominal, GLS standard errors, $\overline{\text{se}}\left(\hat{\beta}_2\right)$, in row (16) is a bit too small. On average the usual GLS standard error understates the true sampling variation of the GLS estimator. However, using the heteroskedasticity robust standard error, HCE1, in row (17), on average closely measures the actual variation $\text{sd}\left(\hat{\beta}_2\right)$.
5. How well does the incorrect GLS estimator do relative to OLS and the correct FGLS estimator? When the random errors are homoskedastic, $\alpha_2 = 0$, the standard deviation of the GLS estimator is larger than that of the OLS estimator. Using GLS when OLS is appropriate is not a good idea. Note that FGLS does almost as well as OLS in this case, so there is not as much of a penalty when the pattern of heteroskedasticity is estimated. When heteroskedasticity is present the incorrect, but reasonable, GLS transformation yields estimates that are more precise than the OLS estimates. In row (15) we see that the ratio $\text{sd}\left(\hat{\beta}_2\right)/\text{sd}\left(b_2\right) < 1$ when $\alpha_2 \neq 0$. Partially curing the heteroskedasticity has produced an improvement. However, the GLS estimator improvement is not as great as for the FGLS estimator when heteroskedasticity is severe, $\alpha_2 = 0.5$.
6. How well does the NR^2 test do in detecting heteroskedasticity? Using the OLS residuals, the rejection rates of the test are $\text{rej}\left(NR^2\right)$ in row (7). When errors are homoskedastic, $\alpha_2 = 0$, the test rejects about 5% of the time as desired. When heteroskedasticity is present the test rejects homoskedasticity a very large percentage of the time, which is also desirable.
7. Finally, compare experiment (4) to experiment (3). These experiments have the same data generating process, except in experiment (3) we have 100 observations in a sample and in experiment (4) we have 5000 observations per sample. With 100 observations the standard deviation of the OLS estimates, which is the true sampling variation, is about 0.36. Using a two standard deviation rule, would being within ± 0.72 of the true parameter value $\beta_2 = 1.0$ be adequately informative for your work? If not, then the sampling variation can be reduced using FGLS, in this case so that the margin of error is ± 0.50. If that is not adequate you will need to build a better model or obtain more sample data. With 5000 observations the two standard deviation margin of error of the OLS estimates is about ± 0.10. Would that be adequate for your work? If so then nothing beyond OLS estimation with robust standard errors is needed. If not, then pursuing FGLS can reduce the margin of error to about ± 0.06. Having more good data facilitates statistical inference.

Experiment (5) is based on a different skedasticity function, $h\left(\mathbf{x}_i\right) = 3u_i/\bar{h}$, where $u_i \sim uniform(1, 11)$ is a uniform random variable, varying over the range (1,11). In this case $\text{var}\left(e_i\right) = h\left(\mathbf{x}_i\right) z_i = \sigma_i^2$ is different for each observation, heteroskedasticity is present, but the variance changes randomly from one observation to the next with no pattern and no relationship to the model explanatory variables or any other variables. This is **unconditional heteroskedasticity** and it has no effect on the properties of the OLS estimator and OLS is the best linear unbiased estimator. The NR^2 test has no ability to detect this type of heteroskedasticity.

CHAPTER 9

Regression with Time-Series Data: Stationary Variables

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain why lags are important in models that use time-series data, and the ways in which lags can be included in dynamic econometric models.
2. Explain what is meant by a serially correlated time series and how we measure serial correlation.
3. Compute the autocorrelations for a time series, graph the corresponding correlogram, and use it to test for serial correlation.
4. Explain the nature of regressions that involve lagged variables and the number of observations that are available.
5. Use autoregressive (AR) and autoregressive distributed lag (ARDL) models to compute forecasts, standard errors of forecasts, and forecast intervals.
6. Explain the assumptions required for AR and ARDL forecasting.
7. Specify and estimate ARDL models. Use serial correlation checks, significance of coefficients, and model selection criteria to choose lag lengths.
8. Test for Granger causality.
9. Use a correlogram of residuals to test for serially correlated errors.
10. Use a Lagrange multiplier test for serially correlated errors.
11. Explain the differences between time-series models for forecasting and time-series models for policy analysis.
12. Estimate and interpret the estimates from finite and infinite distributed lag models.
13. Compute HAC standard errors for least squares estimates. Explain why they are used.
14. Compute nonlinear least squares and generalized least squares estimates for a model with an AR(1) error.
15. Contrast the exogeneity assumption required for HAC standard errors with that required for estimating an AR(1) error model.
16. Compute delay, interim, and total multipliers for finite and infinite distributed lag models.

17. Test for consistency of least squares in the ARDL representation of an infinite distributed lag model.

18. Contrast the assumptions for a finite distributed lag model with those for an infinite distributed lag model.

KEYWORDS

AR(1) error
ARDL(p, q) model
autocorrelation
autoregressive distributed lags
autoregressive error
autoregressive model
correlogram
delay multiplier
distributed lag weight
dynamic models
exogeneity
finite distributed lag
forecast error
forecast intervals
forecasting
generalized least squares
geometrically declining lag
Granger causality
HAC standard errors
impact multiplier
infinite distributed lag
interim multiplier
lag length
lag operator
lagged dependent variable
LM test
moving average
multiplier analysis
nonlinear least squares
sample autocorrelations
serial correlation
standard error of forecast error
stationarity
total multiplier
$T \times R^2$ form of LM test
weak dependence

9.1 Introduction

When modeling relationships between variables, the nature of the data that have been collected has an important bearing on the appropriate choice of an econometric model. In particular, it is important to distinguish between cross-sectional data (data on a number of economic units at a particular point in time) and time-series data (data collected over time on one particular economic unit). Examples of both types of data were given in Section 1.5. When we say "economic units," we could be referring to individuals, households, firms, geographical regions, countries, or some other entity on which data is collected. Because cross-sectional observations on a number of economic units at a given time are often generated by way of a random sample, they are typically uncorrelated. The level of income observed in the Smiths' household, for example, does not affect, nor is it affected by, the level of income in the Jones's household. On the other hand, time-series observations on a given economic unit, observed over a number of time periods, are likely to be correlated. The level of income observed in the Smiths' household in one year is likely to be related to the level of income in the Smiths' household in the year before. Thus, one feature that distinguishes time-series data from cross-sectional data is the likely correlation between different observations. Our challenges for this chapter include testing for and modeling such correlation.

A second distinguishing feature of time-series data is its natural ordering according to time. With cross-sectional data, there is no particular ordering of the observations that is better or more natural than another. One could shuffle the observations and then proceed with estimation without losing any information. If one shuffles time-series observations, there is a danger of confounding what is their most important distinguishing feature: the possible existence of dynamic–evolving relationships between variables. A dynamic relationship is one in which the change in a variable now has an impact on that same variable, or other variables, in one or more future time periods. For example, it is common for a change in the level of an explanatory variable to have behavioral implications for other variables beyond the time period in which it occurred. The consequences of economic decisions that result in changes in economic variables can last a long time. When the income tax rate is increased, consumers have less disposable income, reducing their expenditures

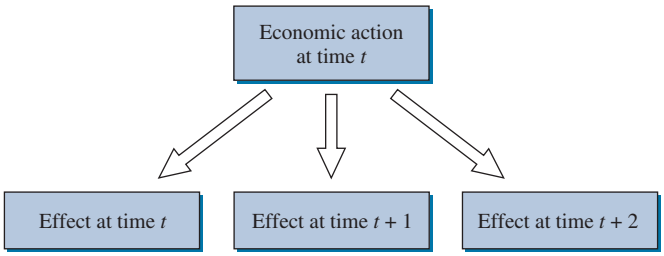


FIGURE 9.1 **The distributed lag effect.**

on goods and services, which reduces profits of suppliers, which reduces the demand for productive inputs, which reduces the profits of the input suppliers, and so on. The effect of the tax increase ripples through the economy. These effects do not occur instantaneously but are spread, or **distributed**, over future time periods. As shown in Figure 9.1, economic actions or decisions taken at one point in time, t, have effects on the economy at time t and also at times $t + 1$, $t + 2$, and so on.

EXAMPLE 9.1 | Plotting the Unemployment Rate and the GDP Growth Rate for the United States

In Figure 9.2(a) and (b), the U.S. quarterly unemployment rate, and the U.S. quarterly growth rate for gross domestic product, from 1948 quarter 1 (1948Q1) to 2016 quarter 1 (2016Q1) are graphed against time. These data can be found in the data file *usmacro*. We wish to understand how series such as these evolve over time, how current values of each data series are correlated with their past values, and how one series might be related to current and past values of another. There are several types of models that can be used to capture the time paths of variables, their correlation structures, and their relationships with the time paths of other variables. Once a model has been selected and estimated, it may be used for **forecasting** future values or for policy analysis. We begin this chapter by describing some of the many possible time-series models and the nature of correlations between current and past values of a data series.

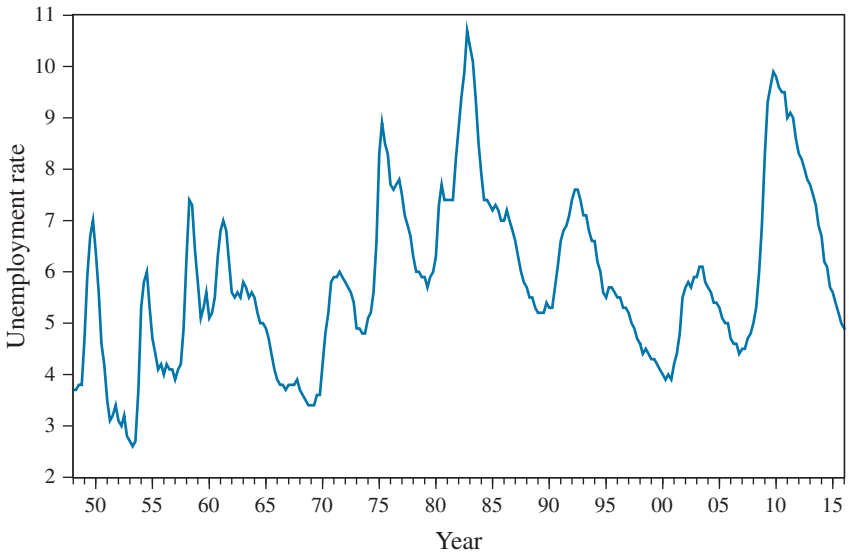


FIGURE 9.2a **U.S. Quarterly unemployment rate 1948Q1 to 2016Q1.**

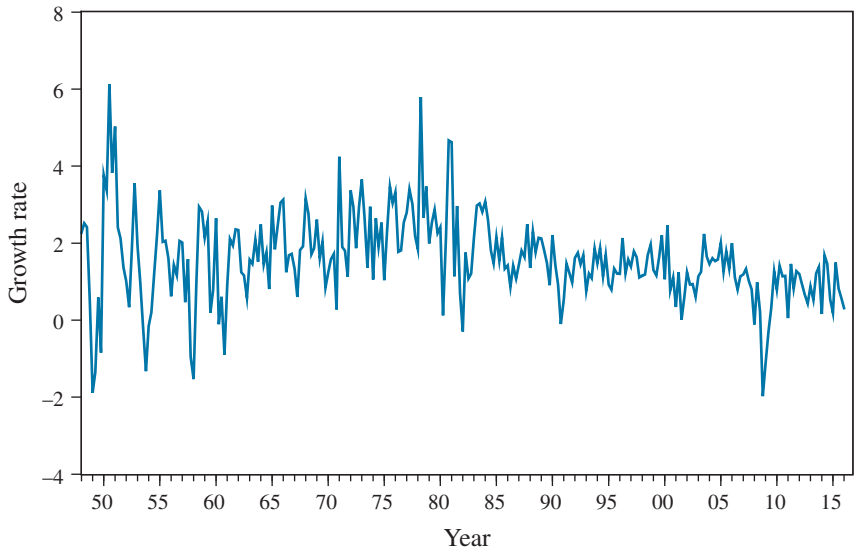


FIGURE 9.2b **U.S. GDP growth rate, 1948Q1 to 2016Q1.**

9.1.1 Modeling Dynamic Relationships

Given that time-series variables are dynamic, in the sense that their current values will be correlated with their past values, and they are related to current and past values of other variables, we need to ask how to model the dynamic nature of relationships. We can do so by introducing lagged variables into the model. These lags can take the form of lagged values of an explanatory variable $(x_{t-1}, x_{t-2}, \ldots, x_{t-q})$, lagged values of a dependent variable $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$, or lagged values of an error term $(e_{t-1}, e_{t-2}, \ldots, e_{t-s})$. In this section, we describe a number of the time-series models that arise from introducing lags of these kinds and explore the relationships between them.

Finite Distributed Lags

Suppose that the value of a variable y depends on current and past values of another variable x, up to q periods into the past. We can write this model as

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} + e_t \tag{9.1}$$

We can think of (y_t, x_t) as denoting the values for y and x in the current period; x_{t-1} means the value of x in the previous period; x_{t-2} is the value of x two periods ago, and so on. Equations like (9.1) might say, for example, that inflation y_t depends not just on the current interest rate x_t but also on the rates in the previous q time periods $x_{t-1}, x_{t-2}, \ldots, x_{t-q}$. Turning this interpretation around as in Figure 9.1, it means that a change in the interest rate now will have an impact on inflation now and in the next q future periods; it takes time for the effect of an interest rate change to fully work its way through the economy. Because of the existence of these lagged effects, equation (9.1) is called a **distributed lag model**. The coefficients β_k are sometimes known as the **lag weights**, and their sequence $\beta_0, \beta_1, \beta_2, \cdots$ is called a **lag pattern**. The model is called a **finite distributed lag** model because the effect of x on y cuts off after a finite number of periods q. Models of this kind can be used for forecasting or policy analysis. In terms of forecasting, we might be interested in using information on past interest rates to forecast future inflation. For policy analysis, a central bank might be interested in how inflation will react now and in the future to a change in the current interest rate.

The notation in (9.1) differs from what we have typically used so far. It is convenient to change the subscript notation on the coefficients: β_s is used to denote the coefficient of x_{t-s} and α is introduced to denote the intercept. Other explanatory variables can be added if relevant, in which case other symbols are needed to denote their coefficients.

Remark

We use many different Greek symbols for regression parameters in this Chapter. Sometimes, it may not seem so, but our goal is clarity.

An Autoregressive Model

An **autoregressive model**, or an **autoregressive process**, is one where a variable y depends on past values of itself. The general representation with p lagged values $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ is called an autoregressive model (process) of order p, abbreviated as AR(p), and is given by

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + e_t \tag{9.2}$$

For example, an AR(2) model for the unemployment rate series U in Figure 9.2(a) would be $U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + e_t$. AR models can be used to describe the time paths of variables and capture their correlations between current and past values; they are generally used for forecasting. Past values are used to forecast future values.

Autoregressive Distributed Lag Models

A more general model that includes both finite distributed lag models and autoregressive models as special cases is the **autoregressive distributed lag** model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{9.3}$$

This model, with p lags of y, the current value x, and q lags of x, is abbreviated as an **ARDL(p, q) model**. The AR component of the name ARDL comes from the regression of y on lagged values of itself; the DL component comes from the distributed lag effect of the lagged x's. For example, an ARDL(2, 1) model relating the unemployment rate U to the growth rate in the economy G would be given by $U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \delta_0 G_t + \delta_1 G_{t-1} + e_t$. ARDL models can be used for both forecasting and policy analysis. Notice that we have used "δ" with no subscript for the intercept and "δ_s" (δ with a subscript) for the coefficient of x_{t-s}. This notation is a little strange, but it avoids introducing another Greek letter for ARDL models.

Infinite Distributed Lag Models

If we take equation (9.1) and assume that the impact of past, lagged x's does not cut off after q periods but goes back into the infinite past, then we have the **infinite distributed lag** (IDL) model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \cdots + e_t \tag{9.4}$$

You might question whether values of x from a long, long time ago would still have an effect on y. You might also wonder how to decide on the cut-off point q for a finite distributed lag. One way out of this dilemma is to assume that the coefficients β_s eventually decline in magnitude with their effect becoming negligible at long lags. There are many possible lag pattern assumptions that could be made to achieve this outcome. To illustrate, consider the **geometrically declining lag** pattern

$$\beta_s = \lambda^s \beta_0, \quad 0 < \lambda < 1, \quad s = 0, 1, 2, \ldots \tag{9.5}$$

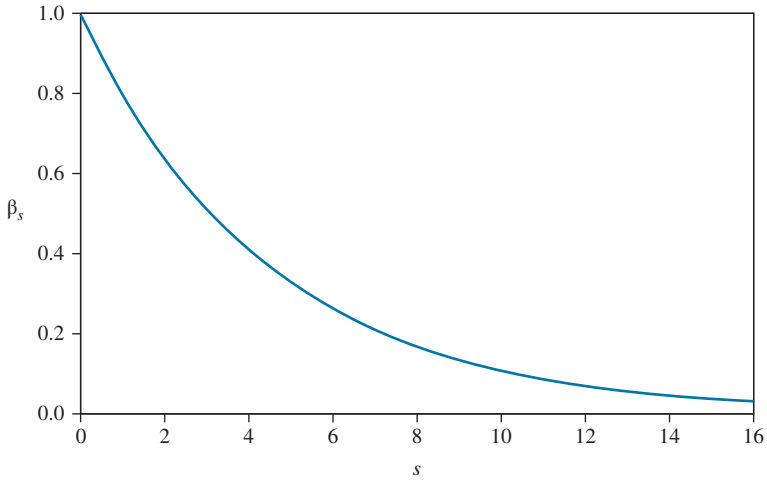


FIGURE 9.3 **Geometrically declining lag pattern.**

A graph of this lag pattern for $\beta_0 = 1$ and $\lambda = 0.8$ is displayed in Figure 9.3. Notice that, as we go back in time (s increases), β_s becomes a smaller and smaller multiple of β_0.

With the assumption in (9.5), we can write

$$y_t = \alpha + \beta_0 x_t + \lambda\beta_0 x_{t-1} + \lambda^2\beta_0 x_{t-2} + \lambda^3\beta_0 x_{t-3} + \cdots + e_t \tag{9.6}$$

Lagging this equation by one period gives the equation for y_{t-1} as

$$y_{t-1} = \alpha + \beta_0 x_{t-1} + \lambda\beta_0 x_{t-2} + \lambda^2\beta_0 x_{t-3} + \lambda^3\beta_0 x_{t-4} + \cdots + e_{t-1}$$

Multiply both sides of this equation by λ to get

$$\lambda y_{t-1} = \alpha\lambda + \lambda\beta_0 x_{t-1} + \lambda^2\beta_0 x_{t-2} + \lambda^3\beta_0 x_{t-3} + \lambda^4\beta_0 x_{t-4} + \cdots + \lambda e_{t-1} \tag{9.7}$$

Subtracting (9.7) from (9.6) gives

$$y_t - \lambda y_{t-1} = \alpha(1 - \lambda) + \beta_0 x_t + e_t - \lambda e_{t-1} \tag{9.8}$$

or

$$y_t = \delta + \theta y_{t-1} + \beta_0 x_t + v_t \tag{9.9}$$

We have made the substitutions $\delta = \alpha(1 - \lambda)$, $\theta = \lambda$, and $v_t = e_t - \lambda e_{t-1}$ so that (9.9) can be recognized as an ARDL model. By making the assumption $\beta_s = \lambda^s\beta_0$, we have been able to turn the IDL model into an ARDL(1, 0) model. On the right-hand side of (9.9), there is one lag of y and the current value of x. We will see later that we can also go in the other direction. More general, ARDL(p, q) models can be turned into more flexible IDL models, providing the lagged coefficients of the IDL eventually decline and become negligible. The ARDL formulation is useful for forecasting; the IDL provides useful information for policy analysis.

An Autoregressive Error Model Another way in which lags can enter a model is through the error term. For example, if the error e_t satisfies the assumptions of an AR(1) model, it can be written as

$$e_t = \rho e_{t-1} + v_t \tag{9.10}$$

with the v_t being uncorrelated. This model means that the random error at time t is related to the random error in the previous time period plus a random component. In contrast to the AR model in (9.2), there is no intercept parameter in (9.10); it is omitted because e_t has a zero mean.

The **AR(1) error** model could be added to any of the models considered so far. To explore one of its implications, suppose that $e_t = \rho e_{t-1} + v_t$ is the error term in the model

$$y_t = \alpha + \beta_0 x_t + e_t \tag{9.11}$$

Substituting $e_t = \rho e_{t-1} + v_t$ into $y_t = \alpha + \beta_0 x_t + e_t$ yields

$$y_t = \alpha + \beta_0 x_t + \rho e_{t-1} + v_t \tag{9.12}$$

From the regression equation (9.11), the error in the previous period, time $t-1$, can be written as

$$e_{t-1} = y_{t-1} - \alpha - \beta_0 x_{t-1} \tag{9.13}$$

Multiplying (9.13) by ρ yields

$$\rho e_{t-1} = \rho y_{t-1} - \rho\alpha - \rho\beta_0 x_{t-1} \tag{9.14}$$

Substituting (9.14) into (9.12) and rearranging yields

$$\begin{aligned} y_t &= \alpha(1-\rho) + \rho y_{t-1} + \beta_0 x_t - \rho\beta_0 x_{t-1} + v_t \\ &= \delta + \theta y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + v_t \end{aligned} \tag{9.15}$$

In the second line of (9.15), we have made the substitutions $\delta = \alpha(1-\rho)$, $\theta = \rho$ and $\beta_1 = -\rho\beta_0$ to show that it is possible to rewrite the AR(1) error model in (9.10) and (9.11) as an ARDL(1, 1) model. Equation (9.15) contains y lagged once, a current value for x, and x lagged once. However, it is a special type of ARDL model because one of its coefficients is equal to the negative product of two of the other coefficients. That is, we have the constraint, or condition, $\beta_1 = -\theta\beta_0$. **Autoregressive error** models with more lags than one can also be transformed to special cases of ARDL models.

Summary and Looking Ahead

We have seen how dynamic relationships between variables can be modeled by including lags in a variety of ways. The various models are summarized in Table 9.1. There is a sense in which most of the models can be viewed as ARDL models or

TABLE 9.1 Summary of Dynamic Models for Stationary Time Series Data

Autoregressive distributed lag model, ARDL(p, q)

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{M1}$$

Finite distributed lag (FDL) model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} + e_t \tag{M2}$$

Infinite distributed lag (IDL) model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \cdots + e_t \tag{M3}$$

Autoregressive model, AR(p)

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + e_t \tag{M4}$$

Infinite distributed lag model with geometrically declining lag weights

$$\beta_s = \lambda^s \beta_0, \quad 0 < \lambda < 1, \quad y_t = \alpha(1-\lambda) + \lambda y_{t-1} + \beta_0 x_t + e_t - \lambda e_{t-1} \tag{M5}$$

Simple regression with AR(1) error

$$y_t = \alpha + \beta_0 x_t + e_t, \quad e_t = \rho e_{t-1} + v_t, \quad y_t = \alpha(1-\rho) + \rho y_{t-1} + \beta_0 x_t - \rho\beta_0 x_{t-1} + v_t \tag{M6}$$

special cases of ARDL models. However, how we interpret and proceed with each model depends on whether the model is to be used for forecasting or policy analysis and on what assumptions are made about the error term in each model. We will examine the various scenarios as we move through the chapter. One pair of assumptions that we make throughout the chapter for all models is that the variables in the models are stationary and weakly dependent. Prior to discussing these two requirements, it is useful to introduce the concept of **autocorrelation** – also known as **serial correlation**.

9.1.2 Autocorrelations

Recall that the concepts of covariance and correlation refer to the degree of linear association between two random variables. If there is no linear association between the variables, then both the covariance and the correlation are zero. When there is some degree of linear association, correlation is the preferred measure because it is unit free and lies within the interval $[-1, 1]$, whereas the magnitude of a covariance will depend on the units of measurement of the two variables. For two random variables, say u and v, their correlation is defined as

$$\rho_{uv} = \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u)\,\text{var}(v)}} \tag{9.16}$$

If u and v are perfectly correlated, then there exist constants c and $d \neq 0$ such that $u = c + dv$, with $\rho_{uv} = 1$ when $d > 0$ and $\rho_{uv} = -1$ when $d < 0$. There is an exact linear relationship. When u and v are uncorrelated, $\rho_{uv} = \text{cov}(u, v) = 0$. Intermediate values of ρ_{uv} measure the degree of linear association.

When dealing with cross-sectional data, it is frequently reasonable to assume that each pair of observations (y_i, x_i) will be uncorrelated with other observations, a characteristic guaranteed by random sampling. In other words, $\text{cov}(y_i, y_j) = 0$ and $\text{cov}(x_i, x_j) = 0$ for $i \neq j$. With time-series data, it is unlikely that these covariances will be zero. If s is close to t, it will almost certainly be the case that $\text{cov}(y_t, y_s) \neq 0$ and $\text{cov}(x_t, x_s) \neq 0$ for $t \neq s$. Glance back at Figure 9.2(a) and (b). If unemployment is higher than average in one quarter, then, in the next quarter, it is more likely to be higher than average again, rather than lower than average. A similar statement can be made for the GDP growth rate. Changes in variables such as unemployment, output growth, inflation, and interest rates are more gradual than abrupt; their values in one period will depend on what happened in the previous period.[1] This dependence means that GDP growth now, for example, will be correlated with GDP growth in the previous period. Successive observations are likely to be correlated. Indeed, in any ARDL model where there is a linear relationship between y_t and its lags, y_t must be correlated with lagged values of itself. Correlations of this kind are called **autocorrelations**. When a variable exhibits correlation over time, we say it is **autocorrelated** or **serially correlated**. We will use these two terms interchangeably.

Let's be more precise about the definition of an autocorrelation. Consider a time series of observations on any variable, $x_1, x_2, \ldots, x_T$, with mean $E(x_t) = \mu_X$ and variance $\text{var}(x_t) = \sigma_X^2$. We assume that μ_X and σ_X^2 do not change over time. The correlation structure between x's that are observed in different time periods is described by the correlation between observations that are one period apart, the correlation between observations that are two periods apart, and so on. If we turn the formula in (9.16) into one that measures the correlation between x_t and x_{t-1}, we have

$$\rho_1 = \frac{\text{cov}(x_t, x_{t-1})}{\sqrt{\text{var}(x_t)\,\text{var}(x_{t-1})}} = \frac{\text{cov}(x_t, x_{t-1})}{\text{var}(x_t)} \tag{9.17}$$

[1] Abrupt changes can occur, particularly with financial data. Models considered in Chapter 14 can accommodate abrupt changes.

The notation ρ_1 is used to denote the population correlation between observations that are one period apart in time, known also as the **population autocorrelation of order 1**. The second equality in (9.17) holds because $\text{var}(x_t) = \text{var}(x_{t-1}) = \sigma_X^2$; we assumed that the variance does not change over time. The population autocorrelation for observations that are s periods apart is

$$\rho_s = \frac{\text{cov}(x_t, x_{t-s})}{\text{var}(x_t)} \qquad s = 1, 2, \ldots \tag{9.18}$$

Sample Autocorrelations Population autocorrelations specified in (9.17) and (9.18) refer to a conceptual time series of observations that goes on forever, starting in the infinite past and continuing into the infinite future, $\cdots, x_{-2}, x_{-1}, x_0, x_1, x_2, \cdots$. **Sample autocorrelations** are obtained using a sample of observations for a finite time period, $x_1, x_2, \ldots, x_T$, to estimate the population autocorrelations. To estimate ρ_1 we use

$$\widehat{\text{cov}}(x_t, x_{t-1}) = \frac{1}{T-1}\sum_{t=2}^{T}(x_t - \bar{x})(x_{t-1} - \bar{x}) \quad \text{and} \quad \widehat{\text{var}}(x_t)\ \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})^2$$

where $\bar{x}$ is the sample mean $\bar{x} = T^{-1}\sum_{t=1}^{T} x_t$. The index of summation in the formula for $\widehat{\text{cov}}(x_t, x_{t-1})$ starts at $t = 2$ because we do not observe x_0. Making the substitutions, and using r_1 to denote the sample autocorrelation at lag 1, we have

$$r_1 = \frac{\sum_{t=2}^{T}(x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{9.19}$$

More generally, **the s-order sample autocorrelation** for a series x, which gives the correlation between observations that are s periods apart (the correlation between x_t and x_{t-s}), is given by

$$r_s = \frac{\sum_{t=s+1}^{T}(x_t - \bar{x})(x_{t-s} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{9.20}$$

This formula is commonly used in the literature and in software and is the one we use to compute autocorrelations in this text, but it is worth mentioning variations of it that are sometimes used. Because $(T - s)$ observations are used to compute the numerator and T observations are used to compute the denominator, an alternative that leads to larger estimates in finite samples is

$$r_s' = \frac{\frac{1}{T-s}\sum_{t=s+1}^{T}(x_t - \bar{x})(x_{t-s} - \bar{x})}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Another modification of (9.20) that has a similar effect is to use only $(T - s)$ observations in the denominator, so that it becomes $\sum_{t=s+1}^{T}(x_t - \bar{x})^2$. Check the computing manuals that go with this book to see which one your software uses.

Testing the Significance of an Autocorrelation It is often useful to test whether a sample autocorrelation is significantly different from zero. That is, a test of $H_0: \rho_s = 0$ against the alternative $H_1: \rho_s \neq 0$. Tests of this nature are useful for constructing models and for checking whether the errors in an equation might be serially correlated. The test statistic for this test is

relatively simple. When the null hypothesis $H_0: \rho_s = 0$ is true, r_s has an approximate normal distribution with mean zero and variance $1/T$. Thus, a suitable test statistic is

$$Z = \frac{r_s - 0}{\sqrt{1/T}} = \sqrt{T} r_s \overset{a}{\sim} N(0, 1) \tag{9.21}$$

The product of the square root of the sample size and the sample autocorrelation r_s has an approximate standard normal distribution. At a 5% significance level, we reject $H_0: \rho_s = 0$ when $\sqrt{T} r_s \geq 1.96$ or $\sqrt{T} r_s \leq -1.96$.

Correlogram A useful device for assessing the significance of autocorrelations is a diagrammatic representation called the **correlogram**. The correlogram, also called the **sample autocorrelation function**, is the sequence of autocorrelations $r_1, r_2, r_3, \ldots$. It shows the correlation between observations that are one period apart, two periods apart, three periods apart, and so on. We indicated that an autocorrelation r_s will be significantly different from zero at a 5% significance level if $\sqrt{T} r_s \geq 1.96$ or if $\sqrt{T} r_s \leq -1.96$. Alternatively, we can say that r_s will be significantly different from zero if $r_s \geq 1.96/\sqrt{T}$ or $r_s \leq -1.96/\sqrt{T}$. A typical diagram for a correlogram will have bars or spikes to represent the magnitudes of the autocorrelations and approximate significant bounds drawn at $\pm 2/\sqrt{T}$, enabling the econometrician to see at a glance which correlations are significant.

EXAMPLE 9.2 | Sample Autocorrelations for Unemployment

Consider the quarterly series for the U.S. unemployment rate found in the data file *usmacro*. It runs from 1948Q1 to 2016Q1, a total of 273 observations. The first four sample autocorrelations for this series, computed from (9.20), are $r_1 = 0.967$, $r_2 = 0.898$, $r_3 = 0.811$, and $r_4 = 0.721$. The value $r_1 = 0.967$ tells us that successive values of unemployment are very highly correlated. With $r_4 = 0.721$, even observations that are four quarters apart are highly correlated. The correlogram for the unemployment rate for the first 24 lags is graphed in Figure 9.4. The heights of the bars represent the correlations. The horizontal line drawn at $2/\sqrt{173} = 0.121$ is the significance bound for positive autocorrelations. Because all the autocorrelations are positive, the negative bound of -0.121 was not included on the graph. The autocorrelations show a gradually declining pattern but remain significantly different from zero until

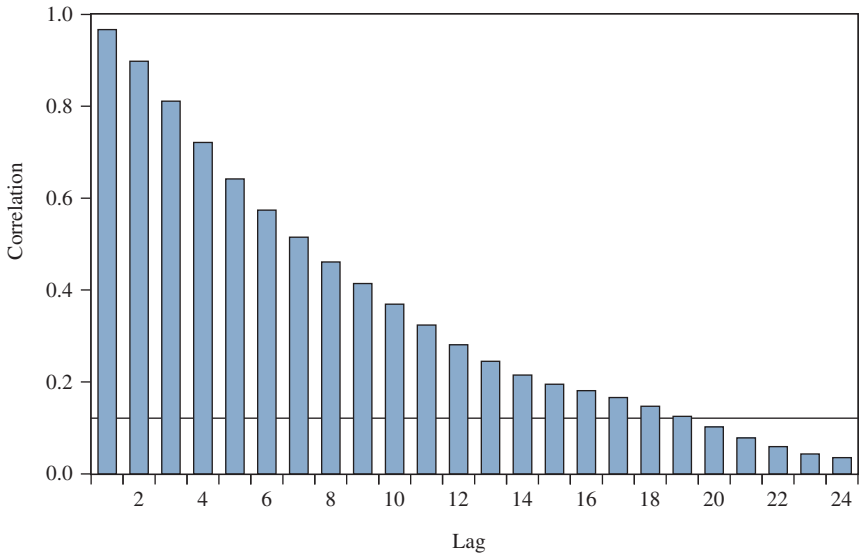


FIGURE 9.4 **Correlogram for U.S. quarterly unemployment rate.**

lag 19, beyond which they are not statistically significant. As the chapter evolves, we will discover that estimates of autocorrelations are important for model construction and checking whether one of our assumptions is violated.

Your software might not produce a correlogram that is exactly the same as Figure 9.4. It might have the correlations on the x-axis and the lags on the y-axis. It could use spikes instead of bars to denote the correlations, it might provide a host of additional information, and its significance bounds might be slightly different than ours. Be prepared! Learn to isolate and focus on the information corresponding to that in Figure 9.4 and do not be disturbed if the output is slightly but not substantially different. If the significance bounds are slightly different, it is because they use a different refinement of the large sample approximation $\sqrt{T}r_s \overset{a}{\sim} N(0,1)$.

EXAMPLE 9.3 | Sample Autocorrelations for GDP Growth Rate

As a second example of sample autocorrelations and the associated correlogram, we consider quarterly data for the U.S. GDP growth rate that can also be found in the data file *usmacro*. In this case, the first four sample autocorrelations are $r_1 = 0.507$, $r_2 = 0.369$, $r_3 = 0.149$, and $r_4 = 0.085$; the correlogram for up to 48 lags is presented in Figure 9.5. These correlations are much smaller than those for the unemployment series, but there is a seemingly strange pattern where the correlations, although not large, oscillate between significance and insignificance at longer lags. This is a complex structure, perhaps attributable to the business cycle.

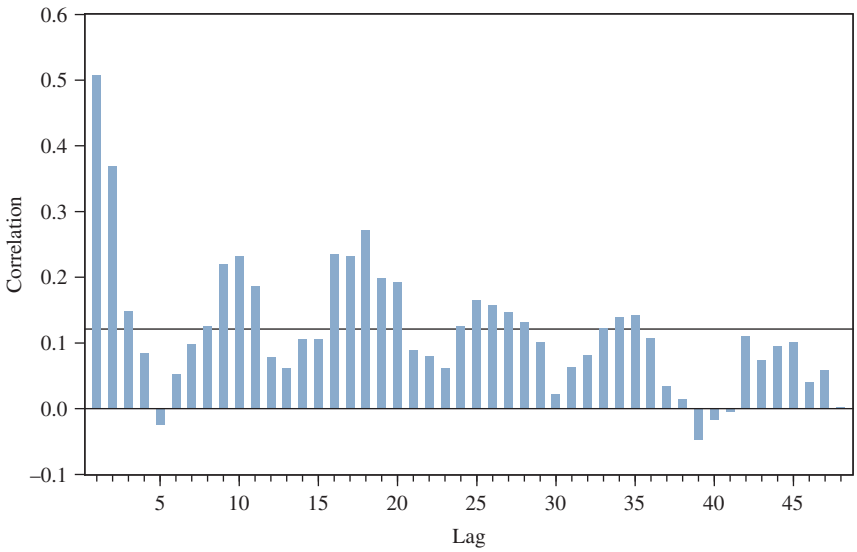


FIGURE 9.5 **Correlogram for growth rate in U.S. GDP.**

9.2 Stationarity and Weak Dependence

A critical assumption that is maintained throughout this chapter is that the variables in our equations are **stationary**. Stationary variables have means and variances that do not change over time and autocorrelations that depend only on how far apart the observations are in time, not on a particular point in time. Specifically, the autocorrelations in (9.18) depend on the time between the periods s, but not the actual point in time t. Implicit in the discussion in Section 9.1.2 was that x_t is stationary. Its mean μ_X, variance σ_X^2, and autocorrelations ρ_s were assumed not to be different for different t. In Examples 9.2 and 9.3, autocorrelations for the unemployment and

growth rates were calculated under the assumption that both are stationary. Saying that a series is stationary implies that, if we took different subsets of observations corresponding to different windows of time, and used them for estimation, we would be estimating the same population quantities, the same mean μ, the same variance σ^2, and the same autocorrelations $\rho_1, \rho_2, \rho_3, \ldots$.

The first task when estimating a relationship with time-series data is to plot the observations on the variables, as we did in Figure 9.2(a) and (b), to gain an appreciation of the nature of your data and to see if there is evidence of nonstationarity. In addition, formal tests known as **unit root tests** can be used to detect nonstationarity. These tests and strategies for estimation with nonstationary variables are considered in Chapter 12. Because checking for nonstationarity is an essential first step, some readers may wish to temporarily jump forward to unit root testing in Chapter 12 before returning to our coverage of estimation and forecasting with stationary variables. For the moment, we note that a stationary variable is one that is not explosive, nor is it trending, and nor does it wander aimlessly without returning to its mean. These features can be illustrated with some graphs. Figure 9.6(a–c) contains graphs of simulated observations on three different variables, plotted against time. Plots of this kind are routinely considered when examining time series variables. The variable y that appears in Figure 9.6(a) is considered stationary because it tends to fluctuate around a constant mean without wandering or trending. On the other hand, x and z that appear in Figure 9.6(b) and (c) possess characteristics of nonstationary variables. In Figure 9.6(b), x tends to wander or is "slow turning," while z in Figure 9.4(c) is trending. These concepts will be defined more precisely in Chapter 12. At the present time, the important thing to remember is that this chapter is concerned with modeling and estimating dynamic relationships between stationary variables whose time series have similar characteristics to those of y. That is, they neither "wander," nor "trend."

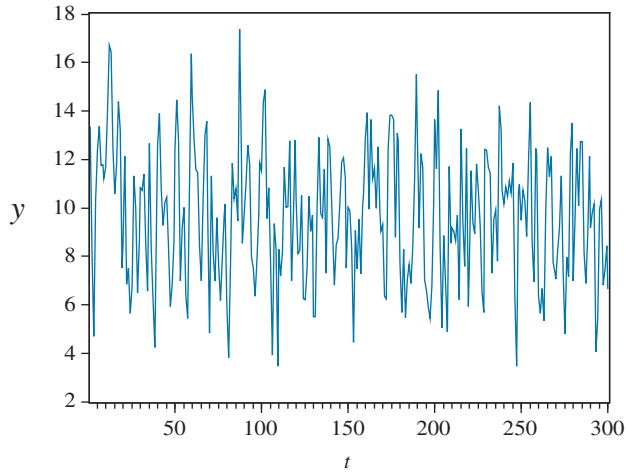
In addition to assuming that the variables are stationary, in this chapter we also assume they are **weakly dependent**. **Weak dependence** implies that, as $s \rightarrow \infty$ (observations get further and further apart in time), they become almost independent. For s large enough, the autocorrelations ρ_s become negligible. When using correlated time-series variables, weak dependence is needed for the least squares estimator to have desirable large sample properties. Typically, stationary variables have weak dependence. However, there are rare exceptions.

EXAMPLE 9.4 | Are the Unemployment and Growth Rate Series Stationary and Weakly Dependent?

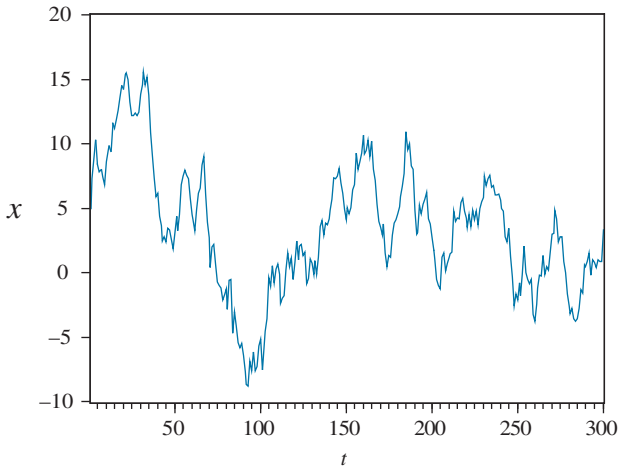
A formal checking of the unemployment and growth rate series for **stationarity** is deferred until unit root tests are introduced in Chapter 12. It is useful, however, to see what tentative conclusions might be drawn from the plots and correlograms of the two series. An examination of the plot for unemployment in Figure 9.2(a) suggests that it has characteristics that make it more similar to Figure 9.6(b) than to Figure 9.6(a). Thus, on the basis of the plot alone, one might be inclined to conclude the unemployment rate is nonstationary. It turns out that a unit root test rejects a null hypothesis of nonstationarity, suggesting that the series can be treated as stationary, but its very high autocorrelations have led to the wandering characteristics exhibited in Figure 9.2(a). Do we have evidence to suggest that the series is weakly dependent? The answer is yes. The autocorrelations in the correlogram in Figure 9.4 are becoming smaller and smaller at longer lags and eventually die out to $r_{24} = 0.035$. Had we considered lags beyond 24, we would find $r_{36} = 0.008$.

Turning to the GDP growth series, we note that its plot in Figure 9.2b has characteristics similar to those of Figure 9.6(a), enabling us to tentatively conclude that it is stationary. GDP growth has ups and downs from one quarter to the next, but it does not keep going up or down for long periods; it returns to the middle, or mean, after a short time. Its correlogram in Figure 9.5 has some significant correlations at long lags, but they are not large and, when autocorrelations beyond those displayed in Figure 9.5 are examined, they die out very quickly, leading us to conclude the series is weakly dependent.

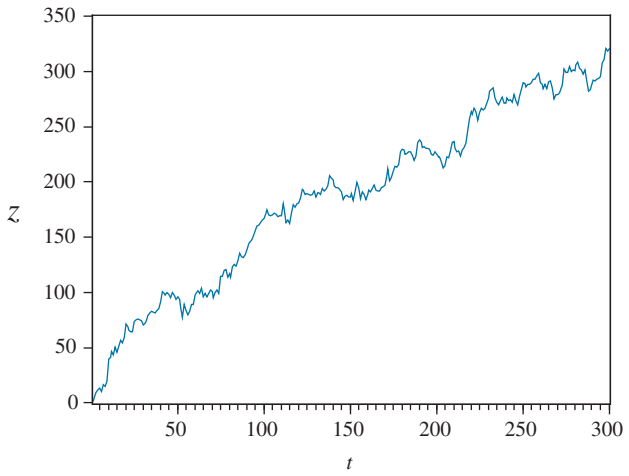
Knowing the unemployment and growth rates are stationary and weakly dependent means that we can proceed to use them for the examples in this chapter devoted to time-series regression models with stationary variables. With the exception of a special case known as cointegration—considered in Chapter 12—variables in time-series regressions must be stationary and weakly dependent for the least squares estimator to be consistent.



(a)



(b)



(c)

FIGURE 9.6 **(a) Time series of a stationary variable; (b) time series of a nonstationary variable that is "slow-turning" or "wandering"; (c) time series of a nonstationary variable that "trends."**

9.3 Forecasting

The forecasting of values of economic variables is a major activity for many institutions including firms, banks, governments, and individuals. Accurate forecasts are important for decision making on government economic policy, investment strategies, the supply of goods to retailers, and a multitude of other things that affect our everyday lives. Because of its importance, you will find that there are whole books and courses that are devoted to the various aspects of forecasting—methods and models for forecasting, ways of evaluating forecasts and their reliability, and practical examples.[2] In this section, we consider forecasting using two different models, an AR model, and an ARDL model. Our focus is on **short-term forecasting**, typically up to three periods into the future.

To introduce the forecasting problem within the context of an ARDL model, suppose that we are given the following ARDL(2,2) model

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \delta_1 x_{t-1} + \delta_2 x_{t-2} + e_t \tag{9.22}$$

Criteria for choosing the numbers of lags for y and x will be discussed in Sections 9.3.3 and 9.4. For the moment, we use two lags of each to describe the essential features of the forecasting problem. A quick comparison of (9.22) with (9.3) reveals a slight difference: the term $\delta_0 x_t$ has been omitted from (9.22). To appreciate why, suppose that we have the sample observations $\{(y_t, x_t), t = 1, 2, \ldots, T\}$ and that we wish to forecast y_{T+1} which, from (9.22), is given by

$$y_{T+1} = \delta + \theta_1 y_T + \theta_2 y_{T-1} + \delta_1 x_T + \delta_2 x_{T-1} + e_{T+1} \tag{9.23}$$

Including $\delta_0 x_t$ in (9.22) would mean including $\delta_0 x_{T+1}$ in (9.23). If the future value x_{T+1} were known, then its inclusion is desirable, but the more likely situation is that both y_{T+1} and x_{T+1} will not be observed at time T when the forecast is made. Thus, dropping x_t in (9.22) is a more practical choice.

Define the information set of all current and past observations on y and x at time t as

$$I_t = \{y_t, y_{t-1}, \ldots, x_t, x_{t-1}, \ldots\} \tag{9.24}$$

Assuming that we are standing at the end of the sample period, having observed y_T and x_T, the one-period ahead forecasting problem is to find a forecast $\hat{y}_{T+1}$ conditional on, or given, the information at time T, $I_T = \{y_T, y_{T-1}, \ldots, x_T, x_{T-1}, \ldots\}$. If the parameters $(\delta, \theta_1, \theta_2, \delta_1, \delta_2)$ are known, the best forecast in the sense that it minimizes conditional mean-squared **forecast error** $E\left[(\hat{y}_{T+1} - y_{T+1})^2 | I_T\right]$ is the conditional expectation $\hat{y}_{T+1} = E(y_{T+1}|I_T)$. We investigate what this implies for the ARDL(2, 2) model in (9.23) and later discuss estimation of the parameters. If we believe that only two lags of y and two lags of x are relevant—they provide the best forecast—we are assuming that

$$\begin{aligned} E(y_{T+1}|I_T) &= E(y_{T+1}|y_T, y_{T-1}, x_T, x_{T-1}) \\ &= \delta + \theta_1 y_T + \theta_2 y_{T-1} + \delta_1 x_T + \delta_2 x_{T-1} \end{aligned} \tag{9.25}$$

Notice the difference between the two conditional expectations: $E(y_{T+1}|I_T)$ conditions on all past observations; $E(y_{T+1}|y_T, y_{T-1}, x_T, x_{T-1})$ conditions on only the two most recent observations. By employing an ARDL(2, 2) model, we are assuming that, for forecasting y_{T+1}, observations from more than two periods in the past do not convey any extra information relative to that contained in the most recent two observations. In addition, for the result in (9.25) to hold, we require

$$E(e_{T+1}|I_T) = 0 \tag{9.26}$$

[2] A comprehensive but relatively advanced treatment is Graham Elliott and Allan Timmermann, *Economic Forecasting*, 2016, Princeton University Press.

For two-period ahead and three-period ahead forecasts, the best forecasts are given respectively by

$$\hat{y}_{T+2} = E(y_{T+2}|I_T) = \delta + \theta_1 E(y_{T+1}|I_T) + \theta_2 y_T + \delta_1 E(x_{T+1}|I_T) + \delta_2 x_T$$
$$\hat{y}_{T+3} = E(y_{T+3}|I_T) = \delta + \theta_1 E(y_{T+2}|I_T) + \theta_2 E(y_{T+1}|I_T) + \delta_1 E(x_{T+2}|I_T) + \delta_2 E(x_{T+1}|I_T)$$

Notice the extra requirements for these two forecasts. We need to know $E(y_{T+2}|I_T)$, $E(y_{T+1}|I_T)$, $E(x_{T+2}|I_T)$ and $E(x_{T+1}|I_T)$. We have estimates of $E(y_{T+2}|I_T)$ and $E(y_{T+1}|I_T)$ readily available from previous periods' forecasts, but $E(x_{T+2}|I_T)$ and $E(x_{T+1}|I_T)$ require extra information. This information can come from independent forecasts or we might be interested in what-if type questions such as if the next two future values of x are $\hat{x}_{T+1}$ and $\hat{x}_{T+2}$, what will be the point and interval forecasts for y_{T+2} and y_{T+3}? If the model is a pure autoregressive one without an x-component, this issue does not arise. In what follows we first consider an example using a pure AR model, and then one with one lagged x. These are both special cases of (9.22). We defer discussion of (9.26) and other assumptions until after the examples.

EXAMPLE 9.5 | Forecasting Unemployment with an AR(2) Model

To demonstrate how to use an AR model for forecasting, we consider the following AR(2) model for forecasting the U.S. unemployment rate U

$$U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + e_t \tag{9.27}$$

The aim is to use observations up to and including 2016Q1 to forecast unemployment in the next three quarters: 2016Q2, 2016Q3, and 2016Q4. The information set at time t is $I_t = \{U_t, U_{t-1}, \ldots\}$. At the time we have observed 2016Q1, it is $I_{2016Q1} = \{U_{2016Q1}, U_{2015Q4}, \ldots\}$. We assume that (9.26) holds which, in general terms for any time period, can be written as $E(e_t|I_{t-1}) = 0$. Past values of unemployment cannot be used to forecast the error in the current period. With this set up, we can write expressions for forecasts for the remainder of 2016 as

$$\hat{U}_{2016Q2} = E(U_{2016Q2}|I_{2016Q1}) = \delta + \theta_1 U_{2016Q1} + \theta_2 U_{2015Q4} \tag{9.28}$$

$$\begin{aligned}\hat{U}_{2016Q3} &= E(U_{2016Q3}|I_{2016Q1}) \\ &= \delta + \theta_1 E(U_{2016Q2}|I_{2016Q1}) + \theta_2 U_{2016Q1}\end{aligned} \tag{9.29}$$

$$\begin{aligned}\hat{U}_{2016Q4} &= E(U_{2016Q4}|I_{2016Q1}) \\ &= \delta + \theta_1 E(U_{2016Q3}|I_{2016Q1}) + \theta_2 E(U_{2016Q2}|I_{2016Q1})\end{aligned} \tag{9.30}$$

Because these expressions all depend on the unknown parameters (δ, θ_1, θ_2), before we can proceed we need to estimate them. We digress for a moment to consider estimation of the AR(2) model.

OLS Estimation of the AR(2) Model for Unemployment

The assumption $E(e_t|I_{t-1}) = 0$ is sufficient for the OLS estimator for $(\delta, \theta_1, \theta_2)$ to be consistent. The OLS estimator will not be unbiased, but consistency gives it a large-sample justification. Assuming that $E(e_t|I_{t-1}) = 0$ is weaker than the strict **exogeneity** assumption. In the general ARDL model, it implies $\text{cov}(e_t, y_{t-s}) = 0$ and $\text{cov}(e_t, x_{t-s}) = 0$ for all $s > 0$ but it does not preclude future values y_{t+s} and x_{t+s}, $s > 0$, from being correlated with e_t. The model in (9.27) can be treated in the same way as the multiple regression model in Chapters 5 and 6, with $U_{t-1} = x_{t1}$ and $U_{t-2} = x_{t2}$. The two lags of the "dependent variable" can be treated as two different explanatory variables. One difference is that the two lags cause us to lose two observations. Instead of having $T = 273$ observations for estimation, only $T - 2 = 271$ are available. From a practical standpoint, this modification is not a concern; the software that you are using will make the necessary adjustments. It is nevertheless useful to fully appreciate how the lagged variables are defined and how their observations enter the estimation procedure. Table 9.2 contains the observations as separate variables in the form they would appear in a spreadsheet. Notice how the observations are lagged and how we lose one observation when U_{t-1} is formed, and two observations when U_{t-2} is formed.

TABLE 9.2 Spreadsheet of Observations for AR(2) Model

t	Quarter	U_t	U_{t-1}	U_{t-2}
1	1948Q1	3.7	•	•
2	1948Q2	3.7	3.7	•
3	1948Q3	3.8	3.7	3.7
4	1948Q4	3.8	3.8	3.7
5	1949Q1	4.7	3.8	3.8
⋮	⋮	⋮	⋮	⋮
271	2015Q3	5.2	5.4	5.6
272	2015Q4	5.0	5.2	5.4
273	2016Q1	4.9	5.0	5.2

Using the observations in Table 9.2 to find OLS estimates of the model in equation (9.27) yields

$$\begin{aligned} \hat{U}_t &= 0.2885 + 1.6128U_{t-1} - 0.6621U_{t-2} \qquad \hat{\sigma} = 0.2947 \\ \text{(se)} &\quad (0.0666)\ (0.0457) \qquad (0.0456) \end{aligned} \tag{9.31}$$

The standard errors in this equation are the conventional least squares standard errors introduced in Chapters 2 and 5. These standard errors and the estimate $\hat{\sigma} = 0.2947$ will be valid with the conditional homoskedasticity assumption $\text{var}(e_t|U_{t-1}, U_{t-2}) = \sigma^2$. In addition, in large samples, the usual t- and F-statistics are valid for testing hypotheses or constructing interval estimates for $(\delta, \theta_1, \theta_2)$. You might wonder whether we need an assumption corresponding to the one made in Chapters 2 and 5, that the errors are serially uncorrelated. It can be shown that one of the assumptions that has already been made, $E(U_t|I_{t-1}) = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2}$, implies that the errors are uncorrelated.[3]

Unemployment Forecasts Having estimated the AR(2) model, we are now in a position to use it for forecasting. Recognizing that the unemployment rates for the two most recent quarters are $U_{2016Q1} = 4.9$ and $U_{2015Q4} = 5$, the forecast for U_{2016Q2} obtained using (9.28) and the estimates in (9.31) is[4]

$$\begin{aligned} \hat{U}_{2016Q2} &= \hat{\delta} + \hat{\theta}_1 U_{2016Q1} + \hat{\theta}_2 U_{2015Q4} \\ &= 0.28852 + 1.61282 \times 4.9 - 0.66209 \times 5 \\ &= 4.8809 \end{aligned} \tag{9.32}$$

Moving to the forecast for two quarters ahead, we have

$$\begin{aligned} \hat{U}_{2016Q3} &= \hat{\delta} + \hat{\theta}_1 \hat{U}_{2016Q2} + \hat{\theta}_2 U_{2016Q1} \\ &= 0.28852 + 1.61282 \times 4.8809 - 0.66209 \times 4.9 \\ &= 4.9163 \end{aligned} \tag{9.33}$$

There is an important difference in the way the forecasts $\hat{U}_{2016Q2}$ and $\hat{U}_{2016Q3}$ are obtained. It is possible to calculate $\hat{U}_{2016Q2}$ using only past observations on U. However, U_{2016Q3} depends on U_{2016Q2}, which is unobserved at time 2016Q1. To overcome this problem, we replace U_{2016Q2} by

[3] See Exercise 9.3 for an example where autocorrelated errors imply an extra lag of the dependent variable should be included.

[4] We carry the coefficient estimates to five decimal places to reduce rounding error.

its forecast $\hat{U}_{2016Q2}$ on the right side of equation (9.33). For forecasting U_{2016Q4}, forecasts for both U_{2016Q3} and U_{2016Q2} are needed on the right side of the equation. Specifically,

$$\begin{aligned}\hat{U}_{2016Q4} &= \hat{\delta} + \hat{\theta}_1 \hat{U}_{2016Q3} + \hat{\theta}_2 \hat{U}_{2016Q2} \\ &= 0.28852 + 1.61282 \times 4.9163 - 0.66209 \times 4.8809 \\ &= 4.986 \end{aligned} \tag{9.34}$$

The forecast unemployment rates for 2016Q2, 2016Q3, and 2016Q4 are approximately 4.88%, 4.92%, and 4.99%, respectively. By the time this book is published, we will be able to compare these forecasts with what actually happened!

9.3.1 Forecast Intervals and Standard Errors

We are typically interested in not just point forecasts but also interval forecasts that give a likely range in which a future value could fall and indicate the reliability of a point forecast. To investigate how to construct a forecast interval, we return to the more general ARDL(2, 2) model

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \delta_1 x_{t-1} + \delta_2 x_{t-2} + e_t$$

and examine the forecast errors for one-period, two-period, and three-period ahead forecasts. The one-period ahead forecast error f_1 is given by

$$\begin{aligned} f_1 = y_{T+1} - \hat{y}_{T+1} &= \left(\delta - \hat{\delta}\right) + \left(\theta_1 - \hat{\theta}_1\right) y_T + \left(\theta_2 - \hat{\theta}_2\right) y_{T-1} + \left(\delta_1 - \hat{\delta}_1\right) x_{T-1} \\ &\quad + \left(\delta_2 - \hat{\delta}_2\right) x_{T-2} + e_{T+1} \end{aligned} \tag{9.35}$$

where $\left(\hat{\delta}, \hat{\theta}_1, \hat{\theta}_2, \hat{\delta}_1, \hat{\delta}_2\right)$ are the least squares estimates. The difference between the forecast $\hat{y}_{T+1}$ and the corresponding realized value y_{T+1} depends on the differences between the actual coefficients and the estimated coefficients and on the value of the random error e_{T+1}. A similar situation arose in Chapters 4 and 6 when we were forecasting using the regression model. What we are going to do differently now is to ignore the error from estimating the coefficients. It is common to do so in time-series forecasting because the variance of the random error is typically large relative to the variances of the estimated coefficients, and the resulting estimator for the forecast error variance retains the property of consistency. This means that we can write the forecast error for one quarter ahead as

$$f_1 = e_{T+1} \tag{9.36}$$

For two periods ahead, the forecast error gets more complicated. In this case, ignoring sampling error from estimating the coefficients, we will be using

$$\hat{y}_{T+2} = \delta + \theta_1 \hat{y}_{T+1} + \theta_2 y_T + \delta_1 \hat{x}_{T+1} + \delta_2 x_T \tag{9.37}$$

to forecast

$$y_{T+2} = \delta + \theta_1 y_{T+1} + \theta_2 y_T + \delta_1 x_{T+1} + \delta_2 x_T + e_{T+2} \tag{9.38}$$

In (9.37), $\hat{y}_{T+1}$ comes from the one-period ahead forecast, but a value for $\hat{x}_{T+1}$ needs to be obtained from elsewhere. To forecast two periods ahead, we will also need $\hat{x}_{T+2}$. These values may come from their own forecasting model, or they might be set by the forecaster to answer what-if type questions. We will assume that these values are given, $\hat{x}_{T+1} = x_{T+1}$ and $\hat{x}_{T+2} = x_{T+2}$, or, alternatively, that we are asking what-if type questions so that we can assume that there is no error from predicting future values of x. Given these assumptions, the two-period ahead forecast error is

$$f_2 = \theta_1 \left(y_{T+1} - \hat{y}_{T+1}\right) + e_{T+2} = \theta_1 f_1 + e_{T+2} = \theta_1 e_{T+1} + e_{T+2} \tag{9.39}$$

For three periods ahead, the error can be shown to be

$$f_3 = \theta_1 f_2 + \theta_2 f_1 + e_{T+3} = \left(\theta_1^2 + \theta_2\right) e_{T+1} + \theta_1 e_{T+2} + e_{T+3} \tag{9.40}$$

Expressing the forecast errors in terms of the e_t's is convenient for deriving expressions for the forecast error variances. With the assumptions $E(e_t|I_{t-1}) = 0$ and $\text{var}(e_t|y_{t-1}, y_{t-2}, x_{t-1}, x_{t-2}) = \sigma^2$, equations (9.36), (9.39), and (9.40) can be used to show that

$$\begin{aligned}\sigma_{f1}^2 &= \text{var}(f_1|I_T) = \sigma^2 \\ \sigma_{f2}^2 &= \text{var}(f_2|I_T) = \sigma^2(1+\theta_1^2) \\ \sigma_{f3}^2 &= \text{var}(f_3|I_T) = \sigma^2\left[(\theta_1^2+\theta_2)^2 + \theta_1^2 + 1\right]\end{aligned} \tag{9.41}$$

The standard errors of the forecast errors are obtained by replacing the unknown parameters in (9.41) by their estimates and then taking the square root. Denoting these standard errors by $\hat{\sigma}_{f1}$, $\hat{\sigma}_{f2}$, and $\hat{\sigma}_{f3}$, $100(1-\alpha)\%$ **forecast intervals** are given by $\hat{y}_{T+j} \pm t_{(1-\alpha/2,\,T-7)}\hat{\sigma}_{fj}$, $j = 1, 2, 3$. The degrees of freedom for the t-distribution are $(T-p-q-1)-2 = T-7$ because five coefficients have been estimated and the two lags have led to a loss of two observations.[5]

EXAMPLE 9.6 | Forecast Intervals for Unemployment from the AR(2) Model

Using the forecast-error variances in (9.41), the estimates in (9.31) and $t_{(0.975,\,268)} = 1.9689$, we can compute the forecast standard errors and 95% forecast intervals presented in Table 9.3. Notice how the forecast standard errors and the widths of the intervals increase as we forecast further into the future, reflecting the extra uncertainty from doing so. It is much harder to be precise about forecasts further into the future. This idea was introduced in Figure 4.2.

TABLE 9.3 **Forecasts and Forecast Intervals for Unemployment from AR(2) Model**

Quarter	Forecast $\hat{U}_{T+j}$	Standard Error of Forecast Error $(\hat{\sigma}_{fj})$	Forecast Interval $\left(\hat{U}_{T+j} \pm 1.9689 \times \hat{\sigma}_{fj}\right)$
2016Q2 ($j = 1$)	4.881	0.2947	(4.301, 5.461)
2016Q3 ($j = 2$)	4.916	0.5593	(3.815, 6.017)
2016Q4 ($j = 3$)	4.986	0.7996	(3.412, 6.560)

EXAMPLE 9.7 | Forecasting Unemployment with an ARDL(2, 1) Model

In this example, we include a lagged value of the growth rate of GDP (G) to see if its inclusion improves the precision of our forecasts. We would expect a high growth rate to lead to less unemployment and a slowdown in the economy to create more unemployment. The least squares estimated model is

$$\begin{aligned}\hat{U}_t &= 0.3616 + 1.5331U_{t-1} - 0.5818U_{t-2} - 0.04824G_{t-1} \\ \text{(se)} \quad &(0.0723) \quad (0.0556) \quad (0.0556) \quad (0.01949) \\ \hat{\sigma} &= 0.2919\end{aligned} \tag{9.42}$$

Apart from the need to supply future values of G necessary for forecasting more than one quarter into the future, the forecasting procedure for an ARDL model is essentially the same as that for a pure AR model. Providing we are content to construct forecast intervals that ignore any error in the specification of future values of G, adding a distributed lag component to the AR model does not require any special treatment. Point and interval forecasts are obtained in the same way. In Exercise 9.16, you are invited to verify the values reported in Table 9.4. For the forecasts for

[5]The large sample distribution theory upon which this forecast interval is based uses a normal distribution rather than a t-distribution. Thus, the interval $\hat{y}_{T+j} \pm z_{1-\alpha/2}\hat{\sigma}_{fj}$ is also used. The t-distribution is frequently chosen in practice to be more conservative.

2016Q3 and 2016Q4, we assumed that $G_{2016Q2} = 0.869$ and $G_{2016Q3} = 1.069$. Comparing the forecasts in Tables 9.3 and 9.4, we find that including the lagged growth rate has increased the point forecasts for unemployment and reduced slightly the standard errors of the forecasts. The main source of the larger point forecasts appears to be the increase in the estimate of the intercept δ from 0.2885 to 0.3616. In addition, although the values $G_{2016Q2} = 0.869$ and $G_{2016Q3} = 1.069$ assume an improved growth rate relative to $G_{2016Q1} = 0.310$, they are still below the sample average growth rate of $\overline{G} = 1.575$.

TABLE 9.4 **Forecasts and Forecast Intervals for Unemployment from ARDL(2, 1) Model**

Quarter	Forecast $\hat{U}_{T+j}$	Standard Error of Forecast Error $(\hat{\sigma}_{fj})$	Forecast Interval $\left(\hat{U}_{T+j} \pm 1.9689 \times \hat{\sigma}_{fj}\right)$
2016Q2 ($j = 1$)	4.950	0.2919	(4.375, 5.525)
2016Q3 ($j = 2$)	5.058	0.5343	(4.006, 6.110)
2016Q4 ($j = 3$)	5.184	0.7430	(3.721, 6.647)

We have considered forecasting with both AR and ARDL models. It remains to point out that forecasting with a finite distributed lag model with no AR component can be carried out within the same framework as forecasting in the linear regression model that we considered in Section 6.4. Instead of the right-hand-side variables being a number of different x's, they comprise a number of lags on the same x.

9.3.2 Assumptions for Forecasting

Throughout this section, we have alluded to the various assumptions that ensure an ARDL model can be estimated consistently and used for forecasting. A summary of these assumptions and some of their implications follows.

F1: The time series y and x are stationary and weakly dependent. How to test this assumption and how to model time series that violate the assumption are considered in Chapter 12.

F2: The conditional expectation $E(y_t|I_{t-1})$ is a linear function of a finite number of lags of y and x. That is,

$$E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} \tag{9.43}$$

where $I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\}$ is defined as the information set at time $t-1$ and represents all past observations at time t. There are a number of things implied by this assumption.

1. Lags of y beyond y_{t-p} and lags of x beyond x_{t-q} do not contribute to the conditional expectation; they cannot improve the forecast of y_t.
2. The error term e_t in the ARDL model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t$$

is such that $E(e_t|I_{t-1}) = 0$.

3. Let $\mathbf{z}_t = (1, y_{t-1}, \ldots, y_{t-p}, x_{t-1}, \ldots x_{t-q})$ denote all right-hand side variables in the ARDL model at time t. The e_t are not serially correlated in the sense that $E(e_t e_s|\mathbf{z}_t, \mathbf{z}_s) = 0$ for $t \neq s$. If the e_t were serially correlated, then at least one

more lag of y should appear in $E(y_t|I_{t-1})$. To gain an intuitive appreciation of why this is so, consider the AR(1) model $y_t = \delta + \theta_1 y_{t-1} + e_t$. Correlation between e_t and e_{t-1} implies that we can write $E(e_t|I_{t-1}) = \rho e_{t-1}$, from which we obtain $E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \rho e_{t-1}$. From the original model, $e_{t-1} = y_{t-1} - \delta - \theta_1 y_{t-2}$, and so

$$\begin{aligned} E(y_t|I_{t-1}) &= \delta + \theta_1 y_{t-1} + \rho(y_{t-1} - \delta - \theta_1 y_{t-2}) \\ &= \delta(1-\rho) + (\theta_1 + \rho) y_{t-1} - \rho\theta_1 y_{t-2} \end{aligned}$$

4. The assumption $E(e_t|I_{t-1}) = 0$ does not preclude feedback from a past error e_{t-j} $(j > 0)$ to current and future values of x. If x is a policy variable whose setting reacts to past values of e and y, the least squares estimator is still consistent and the conditional expectation remains the best forecast. Correlation between e_t and past values of x is excluded, however. If e_t was correlated with x_{t-1} (say), then $E(e_t|I_{t-1}) \neq 0$.

F3: The errors are conditionally homoskedastic, $\text{var}(e_t|\mathbf{z}_t) = \sigma^2$. This assumption is needed for the traditional least squares standard errors to be valid and to compute the forecast standard errors.

9.3.3 Selecting Lag Lengths

So far in our description of an ARDL model and how it can be used for forecasting, we have taken the **lag lengths** p and q as given. A critical assumption to ensure that we had the best forecast in a minimum mean-squared-error sense was that no lags beyond those included in the model contained extra information that could improve the forecast. Technically, this assumption was equivalent to $E(e_t|I_{t-1}) = 0$ where e_t is the equation error term, and $I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\}$ is the set of information prior to period t. A natural question that now arises is: How many lags of y and x should be included? Specifically, in terms of the ARDL(p, q) model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t$$

how do we decide on p and q? There are a number of different criteria that can be used. Because they all do not necessarily lead to the same choice, there is a degree of subjective judgment that must be exercised. It is an area in which econometrics is an art as well as a science.

We can explain three criteria relatively quickly. One is to extend the lag lengths for y and x as long as their estimated coefficients are significantly different from zero. A second is to choose p and q to minimize either the AIC or the SC variable selection criterion. And a third is to evaluate the out-of-sample forecasting performance of each (p, q) combination using a hold-out sample. Testing significance was introduced in Chapter 3 and has been used extensively since. The second and third criteria were discussed in Section 6.4.1. In the remainder of this section, we use the unemployment equation to illustrate how the SC can be used to choose lag lengths.[6] A fourth way of of deciding on p and q is to check for serial correlation in the error term. Since $E(e_t|I_{t-1}) = 0$ implies that the lag lengths p and q are sufficient and the errors are not serially correlated, the presence of serial correlation is an indication we have insufficient lags. Testing for serial correlation is an important topic in its own right, and so we devote Section 9.4 to it.

[6]The SC penalizes additional lags more heavily than does the AIC and hence leads to a more parsimonious model. It is generally preferred to the AIC that can select a model with too many lags even when the sample size is infinitely large. For details, see Russell Davidson and James McKinnon (2004), *Econometric Theory and Methods*, Oxford University Press, p.676–677.

EXAMPLE 9.8 | Choosing Lag Lengths in an ARDL(p, q) Unemployment Equation

Our objective is to use the SC to select the number of lags for U and the number of lags for G in the equation

$$U_t = \delta + \theta_1 U_{t-1} + \cdots + \theta_p U_{t-p} + \delta_1 G_{t-1} + \cdots + \delta_q G_{t-q} + e_t$$

When computing the SC for a number of possible lag lengths, it is important that the same number of observations is used to estimate each model; otherwise, the sum-of-squared-errors component in the SC will not be comparable across models. Since lagging variables leads to a loss of observations, and the number of observations lost depends on the lag length, care must be exercised when selecting the period for estimation. We consider a maximum of eight lags for both U and G and, to ensure comparability, our estimation period is from 1950Q1 to 2016Q1 for *all models*. Up to eight observations are used for the lags on the right-hand side of each equation, and the first sample value for U_t is always 1950Q1, giving a total of 265 observations. The SC values for $p = 1, 2, 4, 6, 8$ and $q = 0, 1, 2, \ldots, 8$ are displayed in Table 9.5.[7] There are p lags of U and q lags of G. The SC values for $p = 3, 5, 7$ were omitted because they were dominated by those for the other values of p and did not convey any extra information. Because the SC values are negative, the minimizing values for p and q are those that lead to the "largest negative" entry, namely $p = 2$ and $q = 0$, suggesting that the ARDL(2, 0) model $U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + e_t$ is suitable. Other things to notice are that the relatively large increases in the SC if U_{t-2} is dropped and that more than two lags of U_t are not favored by the SC irrespective of the value of q.

Since we have also used an ARDL(2, 1) model with G_{t-1} included, we ask whether there is any evidence to support its inclusion. It turns out that, if we go back and start the sample from 1948Q3, dropping the first two observations to accommodate two lags, the SC values for the ARDL(2, 0) and ARDL(2, 1) models are −2.393 and −2.395, respectively. In this case, there is a slight preference for including G_{t-1}. Moreover, as we have seen from equation (9.42), the coefficient of G_{t-1} is significantly different from zero at a 5% significance level. Its p-value for a zero null hypothesis is 0.014.

TABLE 9.5 SC Values for ARDL(p, q) Unemployment Equation

Lag	SC				
q/p	1	2	4	6	8
0	−1.880	−2.414	−2.391	−2.365	−2.331
1	−2.078	−2.408	−2.382	−2.357	−2.323
2	−2.063	−2.390	−2.361	−2.337	−2.302
3	−2.078	−2.407	−2.365	−2.340	−2.306
4	−2.104	−2.403	−2.362	−2.331	−2.297
5	−2.132	−2.392	−2.353	−2.346	−2.312
6	−2.111	−2.385	−2.346	−2.330	−2.292
7	−2.092	−2.364	−2.325	−2.309	−2.271
8	−2.109	−2.368	−2.327	−2.307	−2.269

9.3.4 Testing for Granger Causality

Granger causality[8] refers to the ability of lags of one variable to contribute to the forecast of another variable. In the context of the ARDL model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t$$

we say that x does not "Granger cause" y if

$$E(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x_{t-1}, x_{t-2}, \ldots x_{t-q}) = E(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-p})$$

Thus, testing for Granger causality is equivalent to testing

$$H_0 : \delta_1 = 0, \quad \delta_2 = 0, \ldots, \delta_q = 0$$
$$H_1 : \text{at least one } \delta_i \neq 0$$

[7] The AIC and SC values that are reported are computed using the formulas given in equations (6.43) and (6.44). Your software may provide different values that are based on more general formulas that use a likelihood function. To get the likelihood-based values, you need to add $[1 + \ln(2\pi)] \cong 2.8379$ to the entries in Table 9.4. Adding or subtracting a constant does not change the lag length that minimizes AIC or SC.

[8] Granger, C.W.J. (1969), "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica* 37, 424–38.

It can be performed using the F-test introduced in Chapter 6 for testing joint linear hypotheses. Rejection of H_0 implies x Granger causes y. Note that if x Granger causes y, it does not necessarily imply a direct causal relationship between x and y. It means that having information on past x values will improve the forecast for y. Any causal effect can be an indirect one.

EXAMPLE 9.9 | Does the Growth Rate Granger Cause Unemployment?

To answer this question, we first return to the ARDL(2, 1) model whose estimates were given in equation (9.42). Specifically,

$$\begin{array}{lllll} \hat{U}_t = & 0.3616 + 1.5331U_{t-1} & - 0.5818U_{t-2} & - 0.04824G_{t-1} \\ \text{(se)} & (0.0723) \ (0.0556) & (0.0556) & (0.01949) \end{array}$$

In this model, testing whether G Granger causes U is equivalent to testing the significance of the coefficient of G_{t-1}. It can be carried out with a t- or an F-test. For example, the F-value is

$$F = t^2 = (0.04824/0.01949)^2 = 6.126$$

It exceeds the 5% critical value of $F_{(0.95,\,1,\,267)} = 3.877$, leading us to conclude that G Granger causes U.

To illustrate how the test works when more than one lag is being tested, consider the following model with four lags of G

$$U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \delta_1 G_{t-1} + \delta_2 G_{t-2} + \delta_3 G_{t-3} + \delta_4 G_{t-4} + e_t$$

In this model, testing whether G Granger causes U is equivalent to testing

$$H_0 : \delta_1 = 0, \ \delta_2 = 0, \ \delta_3 = 0, \ \delta_4 = 0$$
$$H_1 : \text{at least one } \delta_i \neq 0$$

The restricted model obtained by assuming that H_0 is true is $U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + e_t$. If we compute an F-value using the restricted and unrestricted sums of squared errors, it is important to make sure that both models use the same number of observations, in this case, 269, for the sample period 1949Q1 to 2016Q1. The F-value for the test is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} = \frac{(23.2471 - 21.3020)/4}{21.3020/(269-7)} = 5.981$$

Because $F = 5.981$ is greater than the 5% critical value $F_{(0.95,\,4,\,262)} = 2.406$, we reject H_0 and conclude that G does Granger cause U.

9.4 Testing for Serially Correlated Errors

Consider again the ARDL(p, q) model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t$$

with $I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\}$ defined as the information set at time $t-1$ and representing all past observations at time t. To keep the notation and exposition relatively simple, suppose $p = q = 1$. One implication of forecasting assumption F2, that all relevant lags have been included in the conditional expectation $E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1}$, is that the errors e_t are serially uncorrelated. For the absence of serial correlation, we require the conditional covariance between any two different errors to be zero. That is, $E(e_t e_s|\mathbf{z}_t, \mathbf{z}_s) = 0$ for all $t \neq s$ where $\mathbf{z}_t = (1, y_{t-1}, x_{t-1})$ denotes all right-hand-side variables in the ARDL model at time t. If $E(e_t e_s|\mathbf{z}_t, \mathbf{z}_s) \neq 0$, then $E(e_t|I_{t-1}) \neq 0$ that, in turn, implies $E(y_t|I_{t-1}) \neq \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1}$. Thus, one way of assessing whether sufficient lags have been included to get the best forecast is to test for serially correlated errors.

Not using the best model for forecasting is not the only implication of serially correlated errors. If $E(e_t e_s|\mathbf{z}_t, \mathbf{z}_s) \neq 0$ for $t \neq s$, then the usual least squares standard errors are invalid. The possibility of invalid standard errors is relevant not just for forecasting equations but also for equations used for policy analysis to be discussed in Section 9.5. For these reasons, testing for serially correlated errors is routine practice when estimating time series regressions. We discuss three tests for this purpose – checking the correlogram of the least squares residuals, a Lagrange multiplier test, and the Durbin–Watson test.

9.4.1 Checking the Correlogram of the Least Squares Residuals

In Section 9.1.2, we saw how the correlogram can be used to examine the nature of the autocorrelations of a time series and to test whether these autocorrelations are significantly different from zero. The autocorrelations for the unemployment and growth series were investigated in Examples 9.2 and 9.3, respectively. In a similar way, we can use the correlogram of the least squares residuals to check for serially correlated errors. Because the errors e_t are unobserved, their correlogram cannot be checked directly. However, we can obtain the least squares residuals $\hat{e}_t = y_t - \hat{\delta} - \hat{\theta}_1 y_{t-1} - \hat{\delta}_1 x_{t-1}$ as estimates of the e_t and examine their autocorrelations. Noting that the mean of the least squares residuals is zero and adapting equation (9.20), we can write the kth order autocorrelation for the residuals as

$$r_k = \frac{\sum_{t=k+1}^{T} \hat{e}_t \hat{e}_{t-k}}{\sum_{t=1}^{T} \hat{e}_t^2} \tag{9.45}$$

Ideally, for the correlogram to suggest no serial correlation, we like to have $|r_k| < 2/\sqrt{T}$ for $k = 1, 2, \ldots$, the 2 being used to approximate 1.96, the critical value for a 5% significance level. However, occasional significant (but small) autocorrelations at long lags do not constitute strong evidence of autocorrelation and are regarded as acceptable.

EXAMPLE 9.10 | Checking the Residual Correlogram for the ARDL(2, 1) Unemployment Equation

For a first example, we return to the ARDL(2,1) model in (9.42), estimated with 271 observations:

$$\begin{aligned} \hat{U}_t &= 0.3616 + 1.5331 U_{t-1} - 0.5818 U_{t-2} - 0.04824 G_{t-1} \\ \text{(se)} \quad & (0.0723) \quad (0.0556) \qquad (0.0556) \qquad (0.01949) \end{aligned}$$

The autocorrelations for its residuals given in the correlogram in Figure 9.7 are generally small and insignificant. There are exceptions at lags 7, 8, and 17, where the autocorrelations exceed the significance bounds. These correlations are at long lags, barely significant, and relatively small $(r_7 = 0.146, r_8 = -0.130, r_{17} = 0.133)$. It is reasonable to conclude that there is no strong evidence of serial correlation.

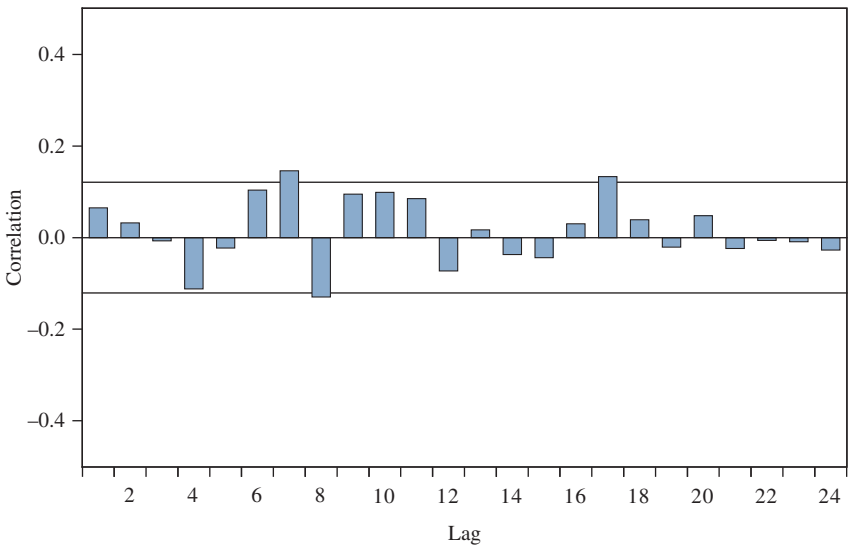


FIGURE 9.7 **Correlogram for residuals from ARDL(2, 1) model.**

EXAMPLE 9.11 | Checking the Residual Correlogram for an ARDL(1, 1) Unemployment Equation

To contrast the outcome in Example 9.10 with one where serial correlation is clearly present, we reestimate the model with U_{t-2} omitted and using 272 observations. If U_{t-2} is an important contributor to the forecasting equation, its omission is likely to lead to serial correlation in the errors. The reestimated equation is

$$\hat{U}_t = 0.4849 + 0.9628U_{t-1} - 0.1672G_{t-1}$$
$$\text{(se)} \quad (0.0842) \quad (0.0128) \qquad (0.0187) \qquad (9.46)$$

and its correlogram is displayed in Figure 9.8. In this case, the first three autocorrelations are significant, and the first two are moderately large $(r_1 = 0.449, r_2 = 0.313)$. We conclude that the errors are serially correlated. More lags are needed to improve the forecasting specification, and the least squares standard errors given in (9.46) are invalid.

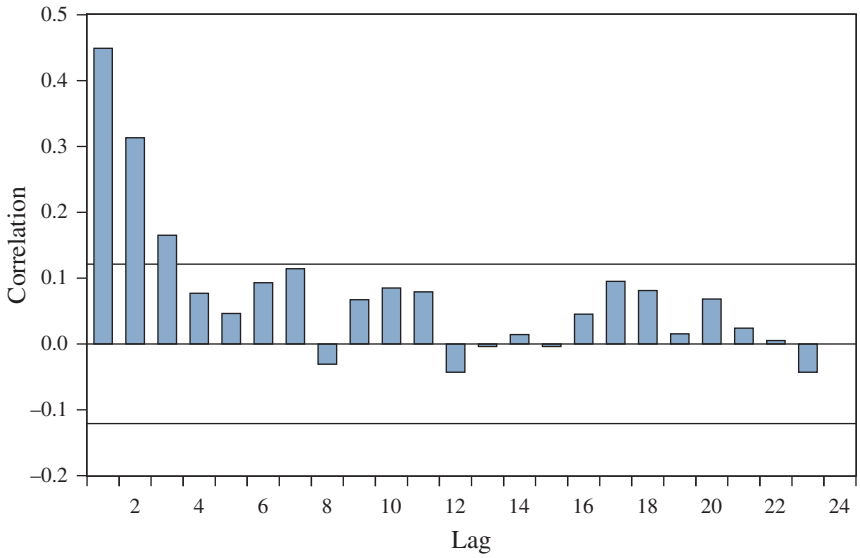


FIGURE 9.8 **Correlogram for residuals from ARDL(1, 1) model.**

9.4.2 Lagrange Multiplier Test

A second test that we consider for testing for serially correlated errors is derived from a general set of hypothesis testing principles that produce Lagrange[9] multiplier (LM) tests. In more advanced courses, you will learn the origin of the term LM. Another example was given in Chapter 8 for testing for heteroskedasticity. The general principle is described in Appendix C.8.4. An advantage of this test is that it readily generalizes to a **joint** test of correlations at more than one lag.

To introduce the test, consider the ARDL(1, 1) model $y_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + e_t$. The null hypothesis for the test is that the errors e_t are uncorrelated. To express this null hypothesis in terms of restrictions on one or more parameters, we can introduce a model for an alternative hypothesis, with that model describing the possible nature of any autocorrelation. We will consider a number of alternative models.

[9]Joseph–Louis Lagrange (1736–1813) was an Italian born mathematician. Statistical tests using the so-called "Lagrange multiplier principle" were introduced into statistics by C.R. Rao in 1948.

Testing for AR(1) Errors In the first instance, we consider an alternative hypothesis that the errors are correlated through the AR(1) process $e_t = \rho e_{t-1} + v_t$ where the new errors v_t satisfy the uncorrelated assumption $\text{cov}(v_t, v_s|\mathbf{z}_t, \mathbf{z}_s) = 0$ for $t \neq s$. In the context of the ARDL(1, 1) model, $\mathbf{z}_t = (1, y_{t-1}, x_{t-1})$. Substituting for e_t in the original equation yields

$$y_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + \rho e_{t-1} + v_t \tag{9.47}$$

Now, if $\rho = 0$, then $e_t = v_t$ and since v_t is not serially correlated, e_t will not be serially correlated. Thus, a test for serial correlation can be set up in terms of the hypotheses $H_0: \rho = 0$ and $H_1: \rho \neq 0$. The obvious way to perform this test if e_{t-1} was observable is to regress y_t on y_{t-1}, $x_{t-1,}$ and e_{t-1} and to then use a t- or F-test to test the significance of the coefficient of e_{t-1}. However, because e_{t-1} is not observable, we replace it by the lagged least squares residuals $\hat{e}_{t-1}$ and then perform the test in the usual way.

Proceeding in this way seems straightforward, but, to complicate matters, applied econometricians have managed to do it in at least four different ways! One of the variations centers around the treatment of the first observation. To appreciate the issue, suppose that we have 100 observations with which to estimate an ARDL(1, 1) model. Because y and x are both lagged once, an effective sample of 99 observations will be used for estimation. There will be 99 residuals $\hat{e}_t$. Replacing e_{t-1} with $\hat{e}_{t-1}$ in (9.47) means that a further observation will be lost leaving 98 for the test equation. An alternative to losing this last observation is to set the initial value of $\hat{e}_{t-1}$ equal to zero so that 99 observations are retained. Doing so is justified because, when H_0 is true, $E(e_{t-1}|\mathbf{z}_{t-1}) = 0$. This is the approach adopted in the automatic commands of the popular software packages Stata and EViews.

The second variation requires a bit more work. As we discovered in Chapter 8, LM tests are such that they can frequently be written as the simple expression $T \times R^2$ where T is the number of sample observations and R^2 is the goodness-of-fit statistic from an auxiliary regression. To derive the relevant auxiliary regression for the autocorrelation LM test, we begin by writing the test equation from (9.47) as

$$y_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + \rho \hat{e}_{t-1} + v_t \tag{9.48}$$

Noting that $y_t = \hat{y}_t + \hat{e}_t = \hat{\delta} + \hat{\theta}_1 y_{t-1} + \hat{\delta}_1 x_{t-1} + \hat{e}_t$, we can rewrite (9.48) as

$$\hat{\delta} + \hat{\theta}_1 y_{t-1} + \hat{\delta}_1 x_{t-1} + \hat{e}_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + \rho \hat{e}_{t-1} + v_t$$

Rearranging this equation yields

$$\begin{aligned} \hat{e}_t &= \left(\delta - \hat{\delta}\right) + \left(\theta_1 - \hat{\theta}_1\right) y_{t-1} + \left(\delta_1 - \hat{\delta}_1\right) x_{t-1} + \rho \hat{e}_{t-1} + v_t \\ &= \gamma_1 + \gamma_2 y_{t-1} + \gamma_3 x_{t-1} + \rho \hat{e}_{t-1} + v_t \end{aligned} \tag{9.49}$$

where $\gamma_1 = \delta - \hat{\delta}$, $\gamma_2 = \theta_1 - \hat{\theta}_1$, and $\gamma_3 = \delta_1 - \hat{\delta}_1$. When testing for autocorrelation by testing the significance of the coefficient of $\hat{e}_{t-1}$, one can estimate (9.48) or (9.49). Both yield the same test result – the same coefficient estimate for $\hat{e}_{t-1}$ and the same t-value. The estimates for the intercept and the coefficients of y_{t-1} and x_{t-1} will be different, however. In (9.49), we are estimating $\left(\delta - \hat{\delta}\right)$, $\left(\theta_1 - \hat{\theta}_1\right)$, and $\left(\delta_1 - \hat{\delta}_1\right)$, instead of δ, θ_1, and δ_1. The auxiliary regression from which the $T \times R^2$ version of the LM test is obtained is (9.49). Because $\left(\delta - \hat{\delta}\right)$, $\left(\theta_1 - \hat{\theta}_1\right)$, and $\left(\delta_1 - \hat{\delta}_1\right)$ are centered around zero, if (9.49) is a regression with significant explanatory power, that power will come from $\hat{e}_{t-1}$.

If $H_0: \rho = 0$ is true, then $\text{LM} = T \times R^2$ has an approximate $\chi^2_{(1)}$ distribution where T and R^2 are the sample size and goodness-of-fit statistic, respectively, from least squares estimation of (9.49). Once again, there are two alternatives depending on whether the first observation is discarded, or $\hat{e}_0$ is set equal to zero.

Testing for MA(1) Errors There are several kinds of models that can be used to try to capture the characteristics of observed sample autocorrelations. These models can be applied to observed time series such as unemployment and the growth rate of GDP or to unobserved errors in a time-series regression model. Up to now only autoregressive models have been discussed. Another useful class of models is what is known as **moving-average** models. You will study these and other models in more depth if you take a time-series course. In Exercise 9.5, you are asked to compare the autocorrelations of an AR(1) model with those of a moving-average model of order one, MA(1). Our task at the moment is to work out a test statistic when an alternative hypothesis of autocorrelation is modeled using the MA(1) process

$$e_t = \phi v_{t-1} + v_t \tag{9.50}$$

The v_t are assumed to be uncorrelated: $\text{cov}\left(v_t, v_s|\mathbf{z}_t, \mathbf{z}_s\right) = 0$ for $t \neq s$. Following the strategy we adopted for the AR(1) error model, combining (9.50) with an ARDL(1,1) model yields

$$y_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + \phi v_{t-1} + v_t \tag{9.51}$$

Notice that $\phi = 0$ implies $e_t = v_t$, and so we can test for autocorrelation through the hypotheses $H_0: \phi = 0$ and $H_1: \phi \neq 0$. Comparing (9.51) with (9.47), we can see that the test for an MA(1) alternative will be exactly the same as the test for an AR(1) alternative providing we can find an estimate $\hat{v}_{t-1}$. Fortunately, we can use the least squares residual $\hat{e}_{t-1}$ to estimate v_{t-1}, just as we did before. That is, $\hat{v}_{t-1} = \hat{e}_{t-1}$. The reason we can do this is that, when H_0 is true, both errors are the same: $e_t = v_t$. Thus, the test for testing against the alternative of MA(1) errors is identical to the test for an alternative of AR(1) errors. The downside of this result is that, when H_0 is rejected, the LM test does not identify which error model is more suitable.

Testing for Higher Order AR or MA Errors The LM test and its variations can be readily extended to alternative hypotheses that are expressed in terms of higher order AR or MA models. For example, suppose that the model for an alternative hypothesis is either an AR(4) or an MA(4) process. Then

$$\begin{aligned} \text{AR(4)}: &\quad e_t = \psi_1 e_{t-1} + \psi_2 e_{t-2} + \psi_3 e_{t-3} + \psi_4 e_{t-4} + v_t \\ \text{MA(4)}: &\quad e_t = \phi_1 v_{t-1} + \phi_2 v_{t-2} + \phi_3 v_{t-3} + \phi_4 v_{t-4} + v_t \end{aligned}$$

The corresponding null and alternative hypotheses for each case are

$$\text{AR(4)} \begin{cases} H_0: \psi_1 = 0, \ \psi_2 = 0, \ \psi_3 = 0, \ \psi_4 = 0 \\ H_1: \text{at least one } \psi_i \text{ is nonzero} \end{cases}$$

$$\text{MA(4)} \begin{cases} H_0: \phi_1 = 0, \ \phi_2 = 0, \ \phi_3 = 0, \ \phi_4 = 0 \\ H_1: \text{at least one } \phi_i \text{ is nonzero} \end{cases}$$

The two alternative test equations corresponding to (9.48) and (9.49) are

$$y_t = \delta + \theta_1 y_{t-1} + \delta_1 x_{t-1} + \psi_1 \hat{e}_{t-1} + \psi_2 \hat{e}_{t-2} + \psi_3 \hat{e}_{t-3} + \psi_4 \hat{e}_{t-4} + v_t \tag{9.52}$$

$$\hat{e}_t = \gamma_1 + \gamma_2 y_{t-1} + \gamma_3 x_{t-1} + \psi_1 \hat{e}_{t-1} + \psi_2 \hat{e}_{t-2} + \psi_3 \hat{e}_{t-3} + \psi_4 \hat{e}_{t-4} + v_t \tag{9.53}$$

We have used the coefficient notation ψ_i from the AR model, but since the test is the same for both AR and MA alternatives, we could equally well have used ϕ_i from the MA model. One can use an F-test to jointly test the significance of the ψ_i in (9.52) or (9.53), or, use the $\text{LM} = T \times R^2$ test computed from (9.53). When H_0 is true, the latter has a $\chi^2_{(4)}$-distribution. Once again, the initial observations can be dropped or set to zero; there will be a slight difference in results from these two alternatives.

EXAMPLE 9.12 | LM Test for Serial Correlation in the ARDL Unemployment Equation

To illustrate the LM test, we apply the $\chi^2 = T \times R^2$ version of the test to the ARDL unemployment equation. Two models are chosen: the ARDL(1, 1) model whose residual correlogram strongly suggested the existence of serially correlated errors and the ARDL(2, 1) model whose correlogram revealed a few small significant correlations, but otherwise was free from serial correlation. Initial values for the $\hat{e}_t$ lost from lagging were set to zero. Table 9.6 contains the test results for AR(k) or MA(k) alternatives for $k = 1, 2, 3, 4$. There is strong evidence that the errors in the ARDL(1, 1) model are serially correlated. With p-values less than 0.0001, the test soundly rejects a null hypothesis of no serial correlation at all four lags. With the ARDL(2,1) model, the results are not so clear cut. At a 5% significance level, a null hypothesis of no serial correlation is not rejected for alternatives with one lag or four lags, but it is rejected for alternatives with two or three lags. Adding a second lag of U_t to the ARDL(1, 1) model has eliminated a large degree of serial correlation in the errors, but some may still remain. In Exercise 9.19, you are invited to test for serial correlation in the errors after adding more lags of U_t and G_t.

TABLE 9.6 LM Test Results for Serial Correlation in the Errors of the Unemployment Equation

Values of k for AR(k) or MA(k)	ARDL(1, 1)		ARDL(2, 1)	
Alternative	Test value	p-Value	Test Value	p-Value
1	66.90	0.0000	2.489	0.1146
2	73.38	0.0000	6.088	0.0476
3	73.38	0.0000	9.253	0.0261
4	73.55	0.0000	9.930	0.0521

9.4.3 Durbin–Watson Test

The sample correlogram and the Lagrange multiplier test are two large-sample tests for serially correlated errors. Their test statistics have their specified distributions in large samples. An alternative test, one that is exact in the sense that its distribution does not rely on a large sample approximation, is the Durbin–Watson test. It was developed in 1950 and, for a long time, was the standard test for $H_0: \rho = 0$ in the AR(1) error model $e_t = \rho e_{t-1} + v_t$. It is used less frequently today because its critical values are not available in all software packages and one has to examine upper and lower critical bounds instead. In addition, unlike the LM and correlogram tests, its distribution no longer holds when the equation contains a lagged dependent variable. A quick rule of thumb, useful when checking your computer output, is that a Durbin–Watson statistic value near 2.0 is compatible with the hypothesis of no serial correlation. Details are provided in Appendix 9A.

9.5 Time-Series Regressions for Policy Analysis

In Section 9.3, we focused on specification, estimation, and use of time-series regressions for forecasting. The main concern was how to use an estimate of an AR conditional expectation

$$E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p}$$

or an ARDL conditional expectation

$$E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q}$$

to forecast the future values $y_{T+1}, y_{T+2}, \ldots$, given the information available at the end of the sample period, I_T. In the AR model, the information set was $I_T = \{y_T, y_{T-1}, y_{T-2}, \ldots\}$; for the ARDL model it was $I_T = \{y_T, x_T, y_{T-1}, x_{T-1}, y_{T-2}, x_{T-2}, \ldots\}$. We were not concerned with the interpretation of individual coefficients, and, providing an adequate number of lags of y (or y

and x) was included in the relevant conditional expectation, we were not concerned with omitted variables. Valid forecasts could be obtained from either of the models or one that contains other explanatory variables and their lags. Moreover, because we were using past data to forecast the future, a current value of x was not included in the ARDL model.

Models for policy analysis differ in a number of ways. The individual coefficients are of interest because they might have a causal interpretation, telling us how much the average outcome of a dependent variable changes when an explanatory variable and its lags change. For example, central banks who set interest rates are concerned with how a change in the interest rate will affect inflation, unemployment, and GDP growth, now and in the future. Because we are interested in the current effect of a change, as well as future effects, the current value of explanatory variables can appear in distributed lag or ARDL models. In addition, omitted variables can be a problem if they are correlated with the included variables because then the coefficients may not reflect causal effects.

Interpreting a coefficient as the change in a dependent variable *caused* by a change in an explanatory variable is in line with the emphasis in Chapters 2–8. With the exception of Section 6.3.1, where we discussed the difference between predictive and causal models, and some sections devoted to prediction, the focus in those chapters was on estimating $\beta_k = \partial E(y_t|\mathbf{x}_t)/\partial x_{tk}$ in the model

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + e_t$$

and on how the interpretation of the β_k changes if one or more variables is expressed in terms of logarithms or if there is some other nonlinear relationship between y_t and x_{tk}. Results from these earlier chapters on the estimation of causal effects also hold for time series regressions providing some critical assumptions hold. Under assumptions MR1–MR5 described in Chapter 5, least squares estimates of the β_k are best linear unbiased. However, there are two of these assumptions that can be very restrictive when working within a time-series framework. Recalling that $\mathbf{X}$ is used to denote all observations in all time periods for the right-hand-side variables, those two assumptions are strict exogeneity, $E(e_t|\mathbf{X}) = 0$, and the absence of serial correlation in the errors, $\text{cov}(e_t, e_s|\mathbf{X}) = 0$ for $t \neq s$. Strict exogeneity implies that there are no lagged dependent variables on the right-hand side, ruling out ARDL models. It also means that the errors are uncorrelated with future x values, an assumption that would be violated if x was a policy variable, such as the interest rate, whose setting was influenced by past values of y, such as the inflation rate. The absence of serial correlation implies that variables omitted from the equation, and whose effect is felt through the error term, must not be serially correlated. Given that time series variables are typically autocorrelated, it is likely to be difficult to satisfy this assumption.

The strict exogeneity assumption can be relaxed if we are content to live with large sample properties. In Section 5.7.3, we noted that the assumptions $E(e_t) = 0$ and $\text{cov}(e_t, x_{tk}) = 0$ for all t and k were sufficient for the least squares estimator to be consistent. Thus, we can still proceed if the errors and right-hand-side variables are contemporaneously uncorrelated, an implication of the lesser assumption of **contemporaneous exogeneity**. In the general framework of an ARDL model, the contemporaneous exogeneity assumption can be written as $E(e_t|\mathbf{z}_t) = 0$ where $\mathbf{z}_t$ denotes all right-hand-side variables that could include both lagged x's and lagged y's. Feedback from current and past y to future x is possible under this assumption, and lagged values of y can be included on the right-hand side. However, as we will discover, for both proper interpretation of coefficients and consistency of estimation, we have to be careful about including the correct number of lags and about the context in which lagged values of y and x arise in the equation. Stronger assumptions often have to be made. In Section 9.1.1, we noted that lagged values of y can arise not just in ARDL models but also in transformations of other models: in a model with an AR(1) error and in an IDL model. The special features of these models are considered in Sections 9.5.3 and 9.5.4. For the OLS standard errors to be valid for large sample inference, the serially uncorrelated error assumption $\text{cov}(e_t, e_s|\mathbf{X}) = 0$ for $t \neq s$ can be weakened to $\text{cov}(e_t, e_s|\mathbf{z}_t, \mathbf{z}_s) = 0$ for $t \neq s$, but we do still need to query whether this assumption is realistic in a time-series setting.

In the following four sections, we are concerned with three main issues that add to our time-series regression results from earlier chapters.

1. Interpretation of coefficients of lagged variables in finite and infinite distributed lag models.
2. Estimation and inference for coefficients when the errors are autocorrelated.
3. The assumptions necessary for interpretation and estimation.

To simplify the discussion, we work with models with only one x and its lags, like those specified at the beginning of this chapter in Table 9.1. Our results and conclusions carry over to models with more than one x and their lags.

9.5.1 Finite Distributed Lags

The finite distributed lag model where we are interested in the impact of current and past values of a variable x on current and future values of a variable y can be written as

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} + e_t \tag{9.54}$$

It is called a *finite* distributed lag because the impact of x on y cuts off after q lags. It is called a *distributed* lag because the impact of a change in x is distributed over future time periods. For the coefficients β_k to represent causal effects, the error term must not be correlated with any omitted variables that are correlated with $\mathbf{x}_t = (x_t, x_{t-1}, \ldots, x_{t-q})$. In particular, since x_t is likely to be autocorrelated, we require e_t not to be correlated with the current and *all past* values of x. This requirement holds if

$$E(e_t | x_t, x_{t-1}, \ldots) = 0 \tag{9.54}$$

It then follows that

$$\begin{aligned} E(y_t | x_t, x_{t-1}, \ldots) &= \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} \\ &= E(y_t | x_t, x_{t-1}, \ldots, x_{t-q}) = E(y_t | \mathbf{x}_t) \end{aligned} \tag{9.55}$$

Once q lags of x have been included in the equation, further lags of x will not have an impact on y.

Given this assumption, a lag-coefficient β_s can be interpreted as the change in $E(y_t|\mathbf{x}_t)$ when x_{t-s} changes by 1 unit, but x is held constant in other periods. Alternatively, if we look forward instead of backward, β_s gives the change in $E(y_{t+s}|\mathbf{x}_t)$ when x_t changes by 1 unit, but x is held constant in other periods. In terms of derivatives

$$\frac{\partial E(y_t|\mathbf{x}_t)}{\partial x_{t-s}} = \frac{\partial E(y_{t+s}|\mathbf{x}_t)}{\partial x_t} = \beta_s \tag{9.56}$$

To further appreciate this interpretation, suppose that x and y have been constant for at least the last q periods and that x_t is increased by 1 unit and then returned to its original level in the next and subsequent periods. Then, using (9.54) but ignoring the error term, the immediate effect will be an increase in y_t by β_0 units. One period later y_{t+1} will increase by β_1 units, then y_{t+2} will increase by β_2 units and so on, up to period $t+q$ when y_{t+q} will increase by β_q units. In period $t+q+1$, the value of y will return to its original level. The effect of a 1-unit change in x_t is **distributed** over the current and next q periods, from which we get the term distributed lag model. The coefficient β_s is called a **distributed-lag weight** or an ***s*-period delay multiplier**. The coefficient $\beta_0\,(s=0)$ is called the **impact multiplier**.

It is also relevant to ask what happens if x_t is increased by 1 unit and then maintained at its new level in subsequent periods $(t+1), (t+2), \cdots$. In this case, the immediate impact will again be β_0; the total effect in period $t+1$ will be $\beta_0 + \beta_1$; in period $t+2$, it will be $\beta_0 + \beta_1 + \beta_2$, and so on. We add together the effects from the changes in all preceding periods. These quantities

are called **interim** or **cumulative multipliers**. For example, the 2-period **interim multiplier** is $(\beta_0 + \beta_1 + \beta_2)$. The **total multiplier** is the final effect on y of the sustained increase after q or more periods have elapsed; it is given by $\sum_{s=0}^{q}\beta_s$.

EXAMPLE 9.13 | Okun's Law

To illustrate the various distributed lag concepts, we introduce an economic model known as Okun's Law.[10] In this model, we again consider a relationship between unemployment and growth of the economy, but we formulate the model differently and use a different data set. Moreover, our purpose is not to forecast unemployment but to investigate the lagged responses of unemployment to growth in the economy. In the basic model for Okun's Law, the change in the unemployment rate from one period to the next depends on the rate of growth of output in the economy:

$$U_t - U_{t-1} = -\gamma(G_t - G_N) \tag{9.57}$$

where U_t is the unemployment rate in period t, G_t is the growth rate of output in period t, and G_N is the "normal" growth rate, which we assume is constant over time. The parameter γ is positive, implying that when the growth of output is above the normal rate, unemployment falls; a growth rate below the normal rate leads to an increase in unemployment. The normal growth rate G_N is the rate of output growth needed to maintain a constant unemployment rate. It is equal to the sum of labor force growth and labor productivity growth. We expect $0 < \gamma < 1$, reflecting that output growth leads to less than one-to-one adjustments in unemployment.

To write (9.57) in the more familiar notation of the multiple regression model, we denote the change in unemployment by $DU_t = \Delta U_t = U_t - U_{t-1}$, we set $\beta_0 = -\gamma$ and $\alpha = \gamma G_N$, and include an error term

$$DU_t = \alpha + \beta_0 G_t + e_t \tag{9.58}$$

Recognizing that changes in output are likely to have a distributed-lag effect on unemployment–not all of the effect will take place instantaneously–we expand (9.58) to include lags of G_t

$$DU_t = \alpha + \beta_0 G_t + \beta_1 G_{t-1} + \beta_2 G_{t-2} + \cdots + \beta_q G_{t-q} + e_t \tag{9.59}$$

To estimate this relationship, we use quarterly Australian data on unemployment and the percentage change in gross domestic product (GDP) from quarter 2, 1978 to quarter 2, 2016. These data are stored in the file *okun5_aus*. The time series for DU and G are graphed in Figure 9.9(a) and (b). There are noticeable jumps in the unemployment rate around 1983, 1992, and 2009; they correspond roughly to periods when there was negative growth but with a lag. At this time, we also note that the series appear to be stationary; tools for more rigorous assessment of stationarity are deferred until Chapter 12.

Least squares estimates of the coefficients and related statistics for equation (9.59) are reported in Table 9.7 for lag

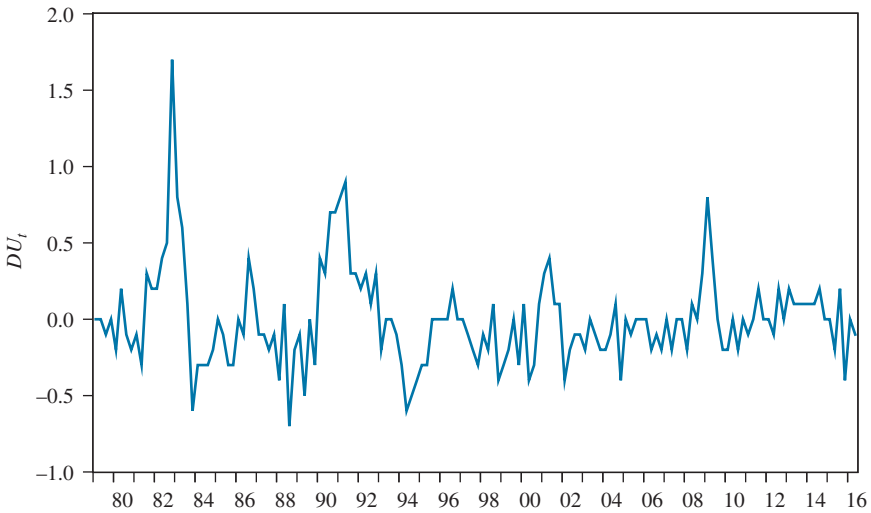


FIGURE 9.9a **Time series for the change in the Australian unemployment rate: 1978Q2 to 2016Q2.**

[10] See O. Blanchard (2009), *Macroeconomics*, 5th edition, Upper Saddle River, NJ, Pearson Prentice Hall, p. 184.

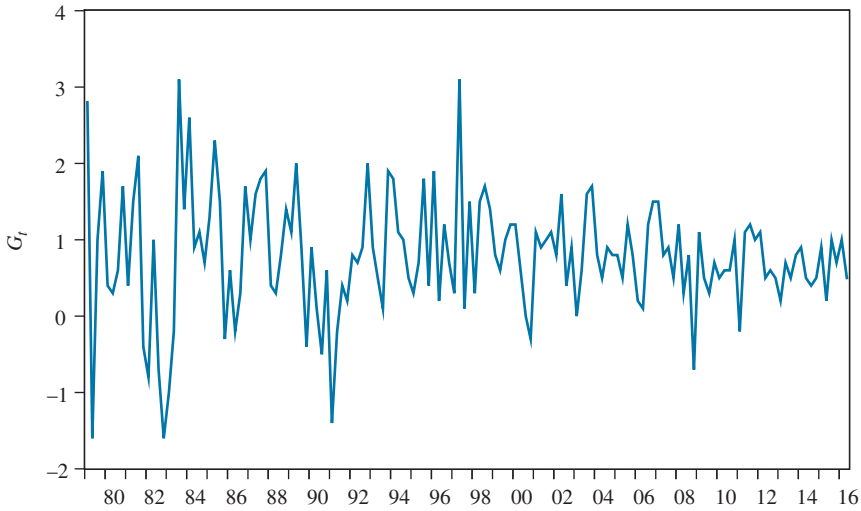


FIGURE 9.9b **Time series for Australian GDP growth: 1978Q2 to 2016Q2.**

lengths $q = 4$ and $q = 5$. All coefficients of G and its lags have the expected negative sign and are significantly different from zero at a 5% significance level, with the exception of that for G_{t-5} when $q = 5$. Given the coefficient of this lag is positive and insignificant, we drop G_{t-5} and settle on a model of order $q = 4$ where all coefficients have the expected negative signs and are significantly different from zero.

What do the estimates for lag length 4 tell us? A 1% increase in the growth rate leads to a fall in the expected unemployment rate of 0.13% in the current quarter, a fall of 0.17% in the next quarter and falls of 0.09%, 0.07%, and 0.06% in two, three, and four quarters from now, respectively. These changes represent the values of the impact multiplier and the one- to four-quarter delay multipliers. The interim multipliers, which give the effect of a sustained increase in the growth rate of 1%, are −0.30 for 1 quarter, −0.40 for 2 quarters, −0.47 for 3 quarters, and −0.53 for 4 quarters. Since we have a lag length of four, −0.53 is also the total multiplier. A summary of these values is presented in Table 9.8. Knowledge of them is important for a government that wishes to keep unemployment below a certain level by influencing the growth rate. If we view γ in equation (9.57) as the total effect of a change in output growth, then its estimate is $\hat{\gamma} = -\sum_{s=0}^{4} b_s = 0.5276$. An estimate of the normal growth rate that is needed to maintain a constant unemployment rate is $\hat{G}_N = \hat{\alpha}/\hat{\gamma} = 0.4100/0.5276 = 0.78\%$ per quarter.

TABLE 9.7 **Estimates for Okun's Law Finite Distributed Lag Model**

Lag Length $q = 5$				
Variable	**Coefficient**	**Standard Error**	***t*-Value**	***p*-Value**
C	0.3930	0.0449	8.746	0.0000
G_t	−0.1287	0.0256	−5.037	0.0000
G_{t-1}	−0.1721	0.0249	−6.912	0.0000
G_{t-2}	−0.0932	0.0241	−3.865	0.0002
G_{t-3}	−0.0726	0.0241	−3.012	0.0031
G_{t-4}	−0.0636	0.0241	−2.644	0.0091
G_{t-5}	0.0232	0.0240	0.966	0.3355
Observations = 148		$R^2 = 0.503$	$\hat{\sigma} = 0.2258$	
Lag Length $q = 4$				
Variable	Coefficient	Standard Error	*t*-Value	*p*-Value
C	0.4100	0.0415	9.867	0.0000
G_t	−0.1310	0.0244	−5.369	0.0000
G_{t-1}	−0.1715	0.0240	−7.161	0.0000
G_{t-2}	−0.0940	0.0240	−3.912	0.0001
G_{t-3}	−0.0700	0.0239	−2.929	0.0041
G_{t-4}	−0.0611	0.0238	−2.563	0.0114
Observations = 149		$R^2 = 0.499$	$\hat{\sigma} = 0.2251$	

TABLE 9.8 **Multipliers for Okun's Law**

Delay Multipliers		Interim Multipliers	
b_0	−0.1310		
b_1	−0.1715	$\sum_{s=0}^{1} b_s$	−0.3025
b_2	−0.0940	$\sum_{s=0}^{2} b_s$	−0.3965
b_3	−0.0700	$\sum_{s=0}^{3} b_s$	−0.4665
b_4	−0.0611	$\sum_{s=0}^{4} b_s$	−0.5276
Total multiplier		$\sum_{s=0}^{4} b_s = -0.5276$	

Assumptions for Finite Distributed Lag Model Before examining some complications that frequently arise with the finite distributed lag model, it is useful to summarize the assumptions that are necessary for OLS estimates to have desirable large sample properties, and the implications of violations of these assumptions. We can also look ahead to what remedies are available to overcome particular violations of assumptions, and their requirements.

FDL1: The time series y and x are stationary and weakly dependent.

FDL2: The finite distributed lag model describing how y responds to current and past values of x can be written as

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_q x_{t-q} + e_t \tag{9.60}$$

FDL3: The error term is exogenous with respect to the current and all past values of x

$$E(e_t | x_t, x_{t-1}, x_{t-2}, \ldots) = 0$$

This assumption ensures

$$E(y_t | x_t, x_{t-1}, x_{t-2}, \ldots) = E(y_t | \mathbf{x}_t)$$

where $\mathbf{x}_t = (x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-q})$. In other words, all relevant lags of x are included in the model. It also implies that there are no omitted variables that are correlated with $\mathbf{x}_t$ and also impact on y_t. This implication raises questions about the Okun's Law example. There are likely to be excluded macro variables that are correlated with GDP growth and that may also impact on the unemployment rate: wage growth, inflation, and interest rates are all possibilities. In the interest of maintaining a relatively simple example, we abstract from these relationships.

FDL4: The error term is not autocorrelated, $\text{cov}(e_t, e_s | \mathbf{x}_t, \mathbf{x}_s) = E(e_t e_s | \mathbf{x}_t, \mathbf{x}_s) = 0$ for $t \neq s$.

FDL5: The error term is homoskedastic, $\text{var}(e_t | \mathbf{x}_t) = E(e_t^2 | \mathbf{x}_t) = \sigma^2$.

Assumptions FDL4 and FDL5 are needed for OLS standard errors, hypothesis tests, and interval estimates to be valid. Since having autocorrelated errors is highly likely, and heteroskedastcity is a possibility, we need to ask how we would proceed when FDL4 and FDL5 are violated. In Chapter 8 when we were faced with the problem of heteroskedastic errors, we considered two possible solutions: (1) using heteroskedasticity consistent robust standard errors for the OLS estimator with no assumptions about the form of the heteroskedasticity being made or (2) making an assumption about the skedastic function and employing a more efficient **generalized least squares** estimator whose standard errors will be valid if the assumption is true. Comparable solutions exist for time series data when FDL4 and FDL5 are violated. It is possible to use the OLS estimator and standard errors known as **HAC (heteroskedasticity and autocorrelation consistent) standard errors**, or **Newey–West standard errors**. Or, we can make some assumption about the nature of the autocorrelation and employ a more efficient generalized squares estimator. In what follows we consider both options. Although the generalized least squares estimator is more efficient, it comes with a cost. In addition to having to make an assumption about the form of the autocorrelation, an exogeneity assumption that is stricter than FDL3 must be made, whereas for OLS with **HAC standard errors**, FDL3 is sufficient.

9.5.2 HAC Standard Errors

To explain the nature of heteroskedasticity and autocorrelation consistent standard errors within a simplified framework, we drop the lagged x's from (9.60), and consider the simple regression model

$$y_t = \alpha + \beta_0 x_t + e_t$$

From Appendix 8A, the least squares estimator for β_0 can be written as

$$b_0 = \beta_0 + \sum_{t=1}^{T} w_t e_t = \beta_0 + \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})\, e_t}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2} = \beta_0 + \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})\, e_t}{s_x^2} \tag{9.61}$$

where s_x^2 is the sample variance for x, using T as the divisor. When e_t was homoskedastic and uncorrelated, we used this result to show that the variance of b_0, conditional on all observations $\mathbf{X}$, is given by (see equation (2.15))

$$\operatorname{var}(b_0|\mathbf{X}) = \frac{\sigma_e^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = \frac{\sigma_e^2}{T s_x^2}$$

For a result that was not conditional on $\mathbf{X}$, we obtained the large sample approximate variance for b_0 from the variance of its asymptotic distribution. This variance is given by $\operatorname{var}(b_0) = \sigma_e^2/T\sigma_x^2$ and uses the fact that s_x^2 is a consistent estimator for σ_x^2. Other terminology is that σ_x^2 is the probability limit of s_x^2, $s_x^2 \xrightarrow{p} \sigma_x^2$ (see Section 5.7, and in particular the discussions following equations (5.34) and (5.35)).

We are now interested in the unconditional variance of b_0 when e_t is both heteroskedastic and autocorrelated. This is a much harder problem. Following similar steps to those sketched out in Section 5.7, we can replace s_x^2 in (9.61) by its probability limit σ_x^2, and $\bar{x}$ by its probability limit μ_x, and then write the large sample variance of b_0 as

$$\begin{aligned}
\operatorname{var}(b_0) &= \operatorname{var}\left(\frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_x)\, e_t}{\sigma_x^2}\right) = \frac{1}{T^2(\sigma_x^2)^2}\operatorname{var}\left(\sum_{t=1}^{T} q_t\right) \\
&= \frac{1}{T^2(\sigma_x^2)^2}\left[\sum_{t=1}^{T}\operatorname{var}(q_t) + 2\sum_{t=1}^{T-1}\sum_{s=1}^{T-t}\operatorname{cov}(q_t, q_{t+s})\right] \\
&= \frac{\sum_{t=1}^{T}\operatorname{var}(q_t)}{T^2(\sigma_x^2)^2}\left[1 + \frac{2\sum_{t=1}^{T-1}\sum_{s=1}^{T-t}\operatorname{cov}(q_t, q_{t+s})}{\sum_{t=1}^{T}\operatorname{var}(q_t)}\right]
\end{aligned} \tag{9.62}$$

where $q_t = (x_t - \mu_x)e_t$. HAC standard errors are obtained by considering estimators for the quantity outside the big brackets and the quantity inside the big brackets. For the quantity outside the brackets, first note that q_t has a zero mean. Then, using $(T-K)^{-1}\sum_{t=1}^{T}\hat{q}_t^2 = (T-K)^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^2\hat{e}_t^2$ as an estimator for $\operatorname{var}(q_t)$, where $\hat{e}_t$ is a least squares residual, $K = 2$ because it is a simple regression, and s_x^2 as an estimator for σ_x^2, an estimator for $\sum_{t=1}^{T}\operatorname{var}(q_t)/T^2(\sigma_x^2)^2$ is given by (see Exercise 9.6)

$$\widehat{\operatorname{var}}_{\mathrm{HCE}}(b_0) = \frac{T\sum_{t=1}^{T}(x_t - \bar{x})^2\hat{e}_t^2}{(T-K)\left(\sum_{t=1}^{T}(x_t - \bar{x})^2\right)^2}$$

Go back and compare this equation with equation (8.9) in Chapter 8. The notation is a little different and the equations are arranged in different ways, but otherwise, they are identical. The quantity outside the brackets in the last line of (9.62) is the large sample unconditional variance of b_0 when there is heteroskedasticity but no autocorrelation. The square root of its estimator $\widehat{\operatorname{var}}_{\mathrm{HCE}}(b_0)$ is the heteroskedasticity consistent, robust standard error. To get a variance estimator for least squares that is consistent in the presence of both heteroskedasticity and autocorrelation, we need to multiply $\widehat{\operatorname{var}}_{\mathrm{HCE}}(b_0)$ by an estimator of the quantity in brackets in (9.62). We will denote this quantity as g.

Several estimators for g have been suggested. To discuss the framework in which they are developed, we simplify g as follows (see Exercise 9.6):

$$g = 1 + \frac{2\sum_{t=1}^{T-1}\sum_{s=1}^{T-t}\text{cov}(q_t, q_{t+s})}{\sum_{t=1}^{T}\text{var}(q_t)} = 1 + \frac{2\sum_{s=1}^{T-1}(T-s)\text{cov}(q_t, q_{t+s})}{T\text{var}(q_t)}$$

$$= 1 + 2\sum_{s=1}^{T-1}\left(\frac{T-s}{T}\right)\tau_s \tag{9.63}$$

where $\tau_s = \text{corr}(q_t, q_{t+s}) = \text{cov}(q_t, q_{t+s})/\text{var}(q_t)$. When there is no serial correlation in the errors, the q_t will also not be autocorrelated, $\tau_s = 0$ for all s, and $g = 1$.To obtain a consistent estimator for g in the presence of autocorrelated errors, the summation in (9.63) is truncated at a lag much smaller than T, the autocorrelations τ_s up to the truncation point are estimated, and the autocorrelations for lags beyond the truncation point are taken as zero. For example, if five autocorrelations are used, the corresponding estimator is

$$\hat{g} = 1 + 2\sum_{s=1}^{5}\left(\frac{6-s}{6}\right)\hat{\tau}_s$$

Alternative estimators differ depending on the number of lags for which the τ_s are estimated and on whether the weights placed on these correlations at each lag are equal to, for example, $(6-s)/6$, or some other alternative. Because there are a large number of possibilities, you will discover that different software packages may yield different HAC standard errors; moreover, different options are possible within a given software package. The message is: Don't be disturbed if you see slightly different HAC standard errors computed for the same problem. Given a suitable estimator $\hat{g}$, the large sample estimator for the variance of b_0, allowing for both heteroscedasticity and autocorrelation in the errors, is

$$\widehat{\text{var}}_{\text{HAC}}(b_0) = \widehat{\text{var}}_{\text{HCE}}(b_0) \times \hat{g}$$

This analysis extends to the finite distributed lag model with q lags and indeed to any time series regression involving stationary variables. The HAC standard errors are given by the square roots of the estimated HAC variances. In Exercise 9.20, you are invited to check whether the errors in the FDL model for Okun's Law in Example 9.13 are autocorrelated and whether using HAC standard errors has an impact on inferences about the multipliers. In Example 9.14 that follows we investigate the impact of serial correlation on the coefficient standard errors for a Phillips curve.

EXAMPLE 9.14 | A Phillips Curve

The Phillips curve has a long history in macroeconomics as a tool for describing the relationship between inflation and unemployment.[11] Our starting point is the model

$$INF_t = INF_t^E - \gamma(U_t - U_{t-1}) \tag{9.64}$$

where INF_t is the inflation rate in period t, INF_t^E denotes inflationary expectations for period t, $DU_t = U_t - U_{t-1}$ denotes the change in the unemployment rate from period $t-1$ to period t, and γ is an unknown positive parameter. It is hypothesized that falling levels of unemployment $(U_t - U_{t-1} < 0)$ reflect excess demand for labor that drives up wages which in turn drives up prices. Conversely, rising levels of unemployment $(U_t - U_{t-1} > 0)$ reflect an excess supply of labor that moderates wage and price increases. The expected inflation rate is included because workers will negotiate wage increases to cover increasing costs from expected inflation, and these wage increases will be transmitted into actual inflation. We assume that inflationary

[11] For a historical review of the development of different versions, see Gordon, R.J. (2008), "The History of the Phillips Curve: An American Perspective", http://nzae.org.nz/wp-content/uploads/2011/08/nr1217302437.pdf, Keynote Address at the Australasian Meetings of the Econometric Society.

expectations are constant over time and set $\alpha = INF_t^E$. In addition, we set $\beta_0 = -\gamma$, and add an error term, in which case the Phillips curve can be written as the simple regression model

$$INF_t = \alpha + \beta_0 DU_t + e_t \tag{9.65}$$

The data used for estimating (9.65) are quarterly Australian data from 1987, Quarter 1 to 2016, Quarter 1, a total of 117 observations, stored in the data file *phillips5_aus*. Inflation is calculated as the percentage change in the Consumer Price Index, with an adjustment in the third quarter of 2000 when Australia introduced a national sales tax. The adjusted time series is graphed in Figure 9.10; the time series for the change in the unemployment rate was previously graphed in Figure 9.9(a). Tests for assessing whether these series are stationary are set as exercises in Chapter 12.

The correlogram of the residuals from least squares estimation of (9.65) is presented in Figure 9.11; approximate 5% significance bounds for the autocorrelations are plotted at $\pm 2/\sqrt{117} = \pm 0.185$. There is evidence of moderate correlations at lags 1–5, and smaller ones at lags 6 and 8. To examine the impact of the autocorrelated errors, in Table 9.9, we report the least squares estimates, and conventional (OLS), HCE and HAC standard errors, *t*-values, and

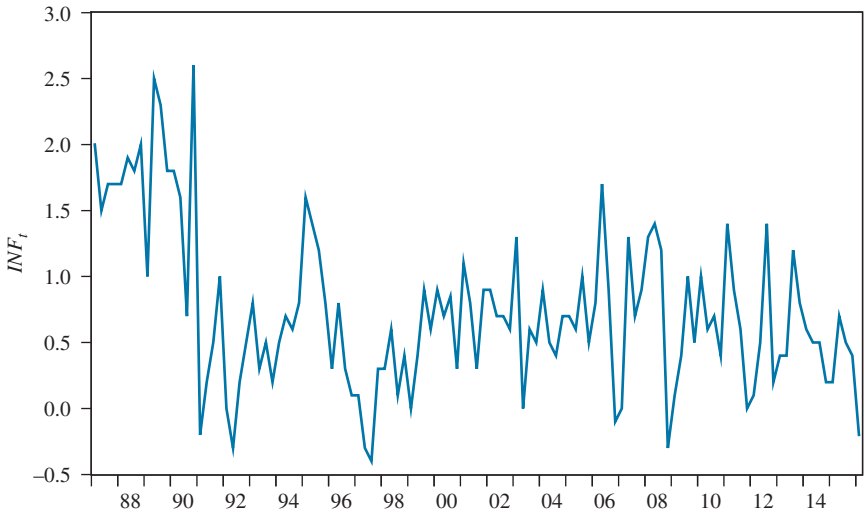


FIGURE 9.10 **Time series for the Australian inflation rate: 1987Q1 to 2016Q1.**

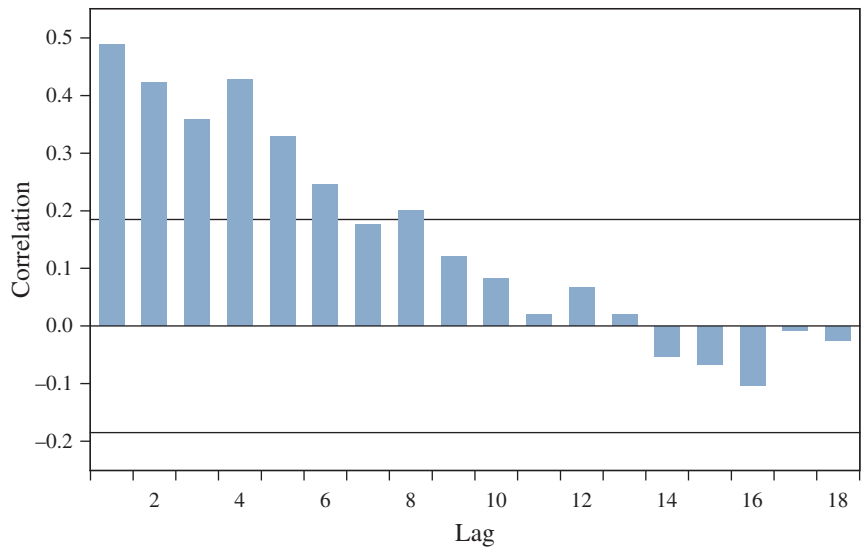


FIGURE 9.11 **Correlogram for least squares residuals from Phillips curve.**

TABLE 9.9 **A Comparison of Conventional (OLS), HCE, and HAC Standard Errors**

Variable	OLS estimate	Standard error			*t*-value			One-tail *p*-value		
		OLS	HCE	HAC	OLS	HCE	HAC	OLS	HCE	HAC
C	0.7317	0.0561	0.0569	0.0915	13.05	12.86	7.99	0.0000	0.0000	0.0000
DU	−0.3987	0.2061	0.2632	0.2878	−1.93	−1.51	−1.39	0.0277	0.0663	0.0844

p-values.[12] The HAC standard errors that allow for autocorrelation and heteroskedasticity are larger than the HCE standard errors that allow only for heteroskedacticity, and the HCE standard errors are larger than the conventional OLS ones that allow for neither heteroskedasticity nor autocorrelation. Thus, ignoring the autocorrelation and heteroskedasticity overstates the reliability of the least squares estimates. Overstating their reliability means that interval estimates will be narrower than they should be and we are more likely to reject a true null hypotheses. Using $t_{(0.975,\,115)} = 1.981$, 95% interval estimates for β_0 are (−0.8070, 0.0096) with conventional standard errors and (−0.9688, 0.1714) with HAC standard errors. With conventional standard errors, a one-tail test and a 5% significance level, we reject $H_0: \beta_2 = 0$. With HCE or HAC standard errors, we do not reject H_0.

9.5.3 Estimation with AR(1) Errors

Using least squares with HAC standard errors overcomes the negative consequences that autocorrelated errors have for least squares standard errors. However, it does not address the issue of finding an estimator that is better in the sense that it has a lower variance. One way to proceed is to make an assumption about the model that generates the autocorrelated errors and to derive an estimator compatible with this assumption. In this section, we examine how to estimate the parameters of the regression model when one such assumption is made, that of AR(1) errors. To keep the exposition free from excessive algebra, we again consider the simple regression model

$$y_t = \alpha + \beta_0 x_t + e_t \tag{9.66}$$

This model can be extended to include extra lags from an FDL model and other variables. The AR(1) error model is given by

$$e_t = \rho e_{t-1} + v_t \qquad |\rho| < 1 \tag{9.67}$$

with the v_t assumed to be uncorrelated random errors with zero mean and constant variances. That is,

$$E(v_t|x_t, x_{t-1}, \ldots) = 0 \qquad \text{var}(v_t|x_t) = \sigma_v^2 \qquad \text{cov}(v_t, v_s|x_t, x_s) = 0 \quad \text{for } t \neq s$$

The assumption $|\rho| < 1$ is required for e_t and y_t to be stationary. From the assumptions about the v_t, we can derive the mean, variance, and autocorrelations for e_t. Conditional on all x's (current, past, and future), it can be shown that e_t has zero mean, constant variance $\sigma_e^2 = \sigma_v^2/(1-\rho^2)$, and autocorrelations $\rho_k = \rho^k$. Thus, the population correlogram that describes the special autocorrelation structure implied by an AR(1) model is $\rho, \rho^2, \rho^3, \ldots$. Because $-1 < \rho < 1$, the AR(1) autocorrelations decline geometrically as the lag increases, eventually becoming negligible. Since there is only one lag of e in the equation $e_t = \rho e_{t-1} + v_t$, you might be surprised to find that autocorrelations at lags greater than one, although declining, are still nonzero.

[12] The HAC standard errors were computed by EViews using a Bartlett kernel, a Newey–West fixed bandwidth of 5, and a degrees of freedom adjustment.

The correlation persists because each e_t depends on all past values of the errors v_t through the equation (see Appendix 9B).[13]

$$e_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots$$

Nonlinear Least Squares Estimation To estimate the AR(1) model described by (9.67) and (9.68), we note, from equation (9.15) in Section 9.1.1, that these equations can be combined and rewritten in the form

$$y_t = \alpha(1-\rho) + \rho y_{t-1} + \beta_0 x_t - \rho\beta_0 x_{t-1} + v_t \tag{9.68}$$

If you are wondering how we get this equation, go back and check out Section 9.1.1. Why is (9.68) useful for estimation? We have transformed the original model in (9.66) with the autocorrelated error term e_t into a new model given by (9.68) that has an error term v_t that is uncorrelated over time. The advantage of doing so is that we can now proceed to find estimates for (α, β_0, ρ) that minimize the sum of squares of uncorrelated errors $S_v = \sum_{t=2}^{T} v_t^2$. The least squares estimator that minimizes the sum of squares of the correlated errors $S_e = \sum_{t=1}^{T} e_t^2$ is not minimum variance and its standard errors are not correct. However, minimizing the sum of squares of uncorrelated errors, S_v, yields an estimator that, in large samples, is best and whose standard errors are correct. Note that this result is in line with earlier practice in the book. The least squares estimator used in Chapters 2 through 7 minimizes a sum of squares of uncorrelated errors.

There is, however, an important distinctive feature about the transformed model in (9.68). Note that the coefficient of x_{t-1} is equal to $-\rho\beta_0$, which is the negative product of ρ (the coefficient of y_{t-1}) and β_0 (the coefficient of x_t). This fact means that, although (9.68) is a linear function of the variables x_t, y_{t-1} and x_{t-1}, it is not a linear function of the parameters (α, β_0, ρ). The usual linear least squares formulas cannot be obtained using calculus to find the values of (α, β_0, ρ) that minimize S_v. Nevertheless, we can still proceed using **nonlinear least squares** to obtain estimates. Nonlinear least squares was introduced in Chapter 6.6. Instead of using formulas to calculate estimates, it uses a numerical procedure to find the estimates that minimize the least squares function.

Generalized Least Squares Estimation To introduce an alternative estimator for (α, β_0, ρ) in the AR(1) error model, we rewrite (9.68) as

$$y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta_0(x_t - \rho x_{t-1}) + v_t \tag{9.69}$$

Defining $y_t^* = y_t - \rho y_{t-1}$, $\alpha^* = \alpha(1-\rho)$ and $x_t^* = x_t - \rho x_{t-1}$, (9.69) becomes

$$y_t^* = \alpha^* + \beta_0 x_t^* + v_t \qquad t = 2, 3, \ldots, T \tag{9.70}$$

If ρ was known, values for the transformed variables y_t^* and x_t^* could be calculated, and least squares applied to (9.70) to find estimates $\hat{\alpha}^*$ and $\hat{\beta}_0$. An estimate for the original intercept is $\hat{\alpha} = \hat{\alpha}^*/(1-\rho)$. This procedure is analogous to that introduced in Section 8.4 where a model with heteroskedastic errors was transformed to one with homoskedastic errors. In that case, the least squares estimator applied to transformed variables y^* and x^* was known as a generalized least squares estimator. Here, we have transformed a model with autocorrelated errors into one with uncorrelated errors. The transformed variables y_t^* and x_t^* are different from those in the heteroscedasticity error case, but, once again, least squares applied to the transformed variables is known as generalized least squares.

Of course, ρ is not known and must be estimated. When the transformed variables are computed using an estimate of ρ, say $\hat{\rho}$, and least squares is applied to these transformed variables, the resulting estimator for α and β_0 is known as a feasible generalized least squares estimator. There are direct parallels with this estimator and the feasible generalized least squares estimator

[13] See Appendix 9B for the derivations.

introduced in Section 8.5. In Section 8.5, parameters in the skedastic function had to be estimated to transform the variables. Here, the parameter in the autocorrelated error model, ρ, needs to be estimated in order to transform the variables.

There are a number of possible estimators for ρ. A simple one is to use r_1 from the sample correlogram. Another one is the least-squares estimate of ρ in a regression of the OLS residuals on their lags. The steps for obtaining the feasible generalized least squares estimator for α and β_0 using this estimator for ρ are as follows:

1. Find least-squares estimates a and b_0 from the equation $y_t = \alpha + \beta_0 x_t + e_t$.
2. Compute the least squares residuals $\hat{e}_t = y_t - a - b_0 x_t$.
3. Estimate ρ by applying least squares to the equation $\hat{e}_t = \rho \hat{e}_{t-1} + \hat{v}_t$. Call this estimate $\hat{\rho}$.
4. Compute values of the transformed variables $y_t^* = y_t - \hat{\rho} y_{t-1}$ and $x_t^* = x_t - \hat{\rho} x_{t-1}$.
5. Apply least squares to the transformed equation $y_t^* = \alpha^* + \beta_0 x_t^* + v_t$.

These steps can also be implemented in an iterative manner. If $\hat{\alpha}$ and $\hat{\beta}_0$ are the estimates obtained in step 5, new residuals can be obtained from $\hat{e}_t = y_t - \hat{\alpha} - \hat{\beta}_0 x_t$, steps 3–5 can be repeated using results from these new residuals, and the process can be continued until the estimates converge. The resulting estimator is often called the **Cochrane–Orcutt** estimator.[14]

Assumptions and Properties Let's pause and take stock of where we are in Section 9.5. In the finite distributed lag model under assumptions FDL1-FDL5, the least squares estimator is consistent, it is minimum variance in large samples, and the usual OLS t-, F-, and χ^2-tests are valid in large samples. However, time-series data are such that assumptions FDL4 (the errors are not autocorrelated) and FDL5 (homoskedasticity), particularly FDL4, might not hold. When FDL4 and FDL5 are violated, the least squares estimator is still consistent, but its usual variance and covariance estimates and standard errors are not correct, leading to invalid t-, F-, and χ^2-tests. One solution to this problem is to use the HAC estimator for variances and covariances and the corresponding HAC standard errors. The least squares estimator is no longer minimum variance when FDL4 and/or FDL5 do not hold, but using HAC variance and covariance estimates means that t-, F-, and χ^2-tests will be valid. Although we examined the use of HAC standard errors in the context of a simple regression model with no lags, they are equally applicable for a finite distributed lag model that includes lags.

A second solution to violation of FDL4 is to assume a specific model for the autocorrelated errors and to use an estimator that is minimum variance for that model. We showed how the parameters of a simple regression model with AR(1) errors can be estimated by (1) nonlinear least squares or (2) feasible generalized least squares. Under two extra conditions, both of these techniques yield a consistent estimator that is minimum variance in large samples, with valid t-, F-, and χ^2-tests. The first extra condition that is needed to achieve these properties is that the AR(1) error model is suitable for modeling the autocorrelated error. We can, however, guard against a failure of this condition using HAC standard errors following nonlinear least squares or feasible generalized least squares estimation. Doing so will ensure t-, F-, and χ^2-tests are valid despite the wrong choice for an autocorrelated error model. The second extra condition is a stronger exogeneity assumption than that in FDL3. To explore this second requirement, consider estimation of α, β_0, and ρ from the nonlinear least squares equation

$$y_t = \alpha(1 - \rho) + \rho y_{t-1} + \beta_0 x_t - \rho \beta_0 x_{t-1} + v_t$$

The exogeneity assumption comparable to FDL3 is

$$E(v_t | x_t, x_{t-1}, x_{t-2}, \ldots) = 0$$

[14] A modification of this process that includes a transformation of the first observation is called the Prais–Winsten estimator. See Exercise 9.7 for details.

Noting that $v_t = e_t - \rho e_{t-1}$, this condition becomes

$$E(e_t - \rho e_{t-1}|x_t, x_{t-1}, x_{t-2}, \ldots) = E(e_t|x_t, x_{t-1}, x_{t-2}, \ldots) - \rho E(e_{t-1}|x_t, x_{t-1}, x_{t-2}, \ldots) = 0$$

Advancing the subscripts in the second term by one period, we can rewrite this condition as

$$E(e_t|x_t, x_{t-1}, x_{t-2}, \ldots) - \rho E(e_t|x_{t+1}, x_t, x_{t-1}, \ldots) = 0$$

For this equation to be true for all possible values of ρ, we require $E(e_t|x_t, x_{t-1}, x_{t-2}, \ldots) = 0$ and $E(e_t|x_{t+1}, x_t, x_{t-1}, \cdots) = 0$. Now, from the law of iterated expectations, $E(e_t|x_{t+1}, x_t, x_{t-1}, \cdots) = 0$ implies $E(e_t|x_t, x_{t-1}, x_{t-2}, \ldots) = 0$. Thus, the exogeneity requirement necessary for nonlinear least squares to be consistent, and it is the same for feasible generalized least squares, is

$$E(e_t|x_{t+1}, x_t, x_{t-1}, \ldots) = 0 \tag{9.71}$$

This requirement implies that e_t and x_{t+1} cannot be correlated. It rules out instances where x_{t+1} is set by a policymaker (such as a central banker setting an interest rate) in response to an error shock in the previous period. Thus, while modeling the autocorrelated error may appear to be a good strategy in terms of improving the efficiency of estimation, it could be at the expense of consistency if the stronger exogeneity assumption is not met. Using least squares with HAC standard errors does not require this stronger assumption.

Modeling of more general forms of autocorrelated errors with more than one lag requires e_t to be uncorrelated with x values further than one period into the future. A stronger exogeneity assumption that accommodates these more general cases and implies (9.71) is the strict exogeneity assumption $E(e_t|\mathbf{X}) = 0$, where $\mathbf{X}$ includes all current, past and future values of the explanatory variables. For general modeling of autocorrelated errors, we replace FDL3 with this assumption.

EXAMPLE 9.15 | The Phillips Curve with AR(1) Errors

In this example, we obtain estimates of the Phillips curve introduced in Example 9.14 under the assumption that its errors can be modeled with an AR(1) process. The data file is *phillips5_aus*. We can, at the outset, conjecture that an AR(1) model might be inadequate. Returning to the correlogram of the least squares residuals in Figure 9.11, the first four sample autocorrelations are $r_1 = 0.489$, $r_2 = 0.358$, $r_3 = 0.422$, and $r_4 = 0.428$. They do not decline exponentially, nor approximately so. Values that start from $r_1 = 0.489$ and decline in line with the properties of an AR(1) model are $r_2 = 0.489^2 = 0.239$, $r_3 = 0.489^3 = 0.117$, and $r_4 = 0.489^4 = 0.057$. Nevertheless, we illustrate the AR(1) error model with this example and later, in Exercise 9.21, explore how we might improve it. Both the nonlinear least squares (NLS) and feasible generalized least squares (FGLS) estimates are reported in Table 9.10, along with the least squares (OLS) estimates and HAC standard errors reproduced from Table 9.9. The NLS and FGLS estimates and their standard errors are almost identical, and the estimates are also similar to those from OLS. The NLS and FGLS standard errors for estimates of β_0 are smaller than the corresponding OLS HAC standard error, perhaps representing an efficiency gain from modeling the autocorrelation. However, one must be cautious with interpretations like this because standard errors are estimates of standard deviations, not the unknown standard deviations themselves.

TABLE 9.10 Phillips Curve Estimates from AR(1) Error Model

Parameter	OLS		NLS		FGLS	
	Estimate	HAC Standard Error	Estimate	Standard Error	Estimate	Standard Error
α	0.7317	0.0915	0.7028	0.0963	0.7029	0.0956
β_0	−0.3987	0.2878	−0.3830	0.2105	−0.3830	0.2087
ρ			0.5001	0.0809	0.4997	0.0799

9.5.4 Infinite Distributed Lags

The finite distributed lag model introduced in Section 9.5.1 assumed that the effect of changes in an explanatory variable x on a dependent variable y cuts off after a finite number of lags q. One way of avoiding the need to specify a value for q is to consider an IDL model where y depends on lags of x that go back into the indefinite past, namely,

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \cdots + e_t \tag{9.72}$$

We introduced this model in Section 9.1.1. For it to be feasible, the β_s coefficients must eventually (but not necessarily immediately) decline in magnitude, becoming negligible at long lags. They have the same multiplier interpretations as in the finite distributed lag case. Specifically,

$$\beta_s = \frac{\partial E(y_t|x_t, x_{t-1}, \ldots)}{\partial x_{t-s}} = s \text{ period delay multiplier}$$

$$\sum_{j=0}^{s} \beta_j = s \text{ period interim multiplier}$$

$$\sum_{j=0}^{\infty} \beta_j = \text{total multiplier}$$

For the total multiplier, we assume the infinite sum converges to a finite value.

Geometrically Declining Lags An obvious disadvantage of the IDL model is its infinite number of parameters. To estimate the lag coefficients in (9.72) with a finite sample of data, some kind of restrictions need to be placed on those coefficients. In Section 9.1.1, we showed that insisting the coefficients decline geometrically through the restrictions $\beta_s = \lambda^s \beta_0$, for $0 < \lambda < 1$, led to the ARDL(1, 0) equation

$$y_t = \delta + \theta y_{t-1} + \beta_0 x_t + v_t \tag{9.73}$$

where $\delta = \alpha(1 - \lambda)$, $\theta = \lambda$, and $v_t = e_t - \lambda e_{t-1}$. Go back and reread Section 9.1.1 to see how (9.73) was derived. By imposing the restrictions, we have been able to reduce the infinite number of parameters to just three. The **delay multipliers** can be calculated from the restrictions $\beta_s = \lambda^s \beta_0$. Using results on the sum of a geometric progression, the interim multipliers are given by

$$\sum_{j=0}^{s} \beta_j = \beta_0 + \beta_0\lambda + \beta_0\lambda^2 + \cdots + \beta_0\lambda^s = \frac{\beta_0(1 - \lambda^{s+1})}{1 - \lambda}$$

and the total multiplier is given by

$$\sum_{j=0}^{\infty} \beta_j = \beta_0 + \beta_0\lambda + \beta_0\lambda^2 + \cdots = \frac{\beta_0}{1 - \lambda}$$

Estimating (9.73) poses some difficulties. If we assume that the original errors e_t are not autocorrelated, then $v_t = e_t - \lambda e_{t-1}$ will be correlated with y_{t-1}, which means $E(v_t|y_{t-1}, x_t) \neq 0$; the least squares estimator will be inconsistent. To see that v_t and y_{t-1} are correlated, note that they both depend on e_{t-1}. It is clear that $v_t = e_t - \lambda e_{t-1}$ depends on e_{t-1}. To see that y_{t-1} also depends on e_{t-1}, we lag (9.72) by one period,

$$y_{t-1} = \alpha + \beta_0 x_{t-1} + \beta_1 x_{t-2} + \beta_2 x_{t-3} + \beta_3 x_{t-4} + \cdots + e_{t-1}$$

Assuming, as we have done in the past, that $E(e_t|x_t, x_{t-1}, x_{t-2}, \ldots) = 0$, meaning we cannot predict e_t given current and past values of x, we have

$$\begin{aligned} E(v_t y_{t-1}|x_{t-1}, x_{t-2}, \ldots) &= E\left[(e_t - \lambda e_{t-1})(\alpha + \beta_0 x_{t-1} + \beta_1 x_{t-2} + \cdots + e_{t-1})\middle| x_{t-1}, x_{t-2}, \ldots\right] \\ &= E\left[(e_t - \lambda e_{t-1})\, e_{t-1}|x_{t-1}, x_{t-2}, \ldots\right] \end{aligned}$$

$$= E\left(e_t e_{t-1}|x_{t-1}, x_{t-2}, \ldots\right) - \lambda E\left(e_{t-1}^2|x_{t-1}, x_{t-2}, \ldots\right)$$
$$= -\lambda \text{var}\left(e_{t-1}|x_{t-1}, x_{t-2}, \ldots\right)$$

where we have used $E\left(e_t e_{t-1}|x_{t-1}, x_{t-2}, \ldots\right) = 0$ from the assumption that e_t and e_{t-1} are conditionally uncorrelated.

One possible consistent estimator for (9.73) is the instrumental variable estimator to be discussed in Chapter 10. It turns out that x_{t-1} is a suitable instrument for y_{t-1}. You are encouraged to think of this as an example when you get to Chapter 10.

There is one special case where least squares applied to (9.73) is a consistent estimator. The inconsistency problem arises because the v_t follow the autocorrelated MA(1) process $v_t = e_t - \lambda e_{t-1}$ **and** y_{t-1} appears on the right side of the equation. The v_t are no longer autocorrelated if the e_t follow the AR(1) process $e_t = \lambda e_{t-1} + u_t$, with the **same** parameter λ, and with the u_t being uncorrelated. In this case, we have

$$v_t = e_t - \lambda e_{t-1} = \lambda e_{t-1} + u_t - \lambda e_{t-1} = u_t$$

Since u_t is not autocorrelated, it will not be correlated with y_{t-1}, and so correlation between y_{t-1} and the error is no longer a source of inconsistency for least squares estimation. Clearly, there is a need to check whether $e_t = \lambda e_{t-1} + u_t$ is a reasonable assumption. A test for this purpose has been proposed by McClain and Wooldridge.[15] Details follow.

Testing for Consistency in the ARDL Representation of an IDL Model

The development of this test starts from the assumption that the errors e_t in the IDL model follow an AR(1) process $e_t = \rho e_{t-1} + u_t$ with parameter ρ that can be different from λ and tests the hypothesis $H_0 : \rho = \lambda$. Under the assumption that ρ and λ are different

$$v_t = e_t - \lambda e_{t-1} = \rho e_{t-1} + u_t - \lambda e_{t-1} = (\rho - \lambda) e_{t-1} + u_t$$

Then, equation (9.73) becomes

$$y_t = \delta + \lambda y_{t-1} + \beta_0 x_t + (\rho - \lambda) e_{t-1} + u_t \tag{9.74}$$

The test is based on whether or not an estimate of the error e_{t-1} adds explanatory power to the regression.

The steps are as follows:

1. Compute the least squares residuals from (9.74) under the assumption that H_0 holds
$$\hat{u}_t = y_t - \left(\hat{\delta} + \hat{\lambda} y_{t-1} + \hat{\beta}_0 x_t\right), \quad t = 2, 3, \ldots, T$$
2. Using the least squares estimate $\hat{\lambda}$ from step 1, and starting with $\hat{e}_1 = 0$, compute recursively $\hat{e}_t = \hat{\lambda}\hat{e}_{t-1} + \hat{u}_t$, $t = 2, 3, \ldots, T$.
3. Find the R^2 from a least squares regression of $\hat{u}_t$ on y_{t-1}, x_t and $\hat{e}_{t-1}$.
4. When H_0 is true, and assuming that u_t is homoskedastic, $(T-1) \times R^2$ has a $\chi^2_{(1)}$ distribution in large samples.

Note that $\hat{u}_t$ can be viewed as equal to y_t after y_{t-1} and x_t have been partialled out. Thus, if the regression in step 3 has significant explanatory power, it will come from $\hat{e}_{t-1}$.

We have described this test in the context of a model with geometrically declining lag weights that leads to an ARDL(1, 0) model with only one lag of y. It can also be performed for ARDL(p, q)

[15] McClain, K.T. and J.M. Wooldridge (1995), "A simple test for the consistency of dynamic linear regression in rational distributed lag models," *Economics Letters*, 48, 235–240.

models where $p > 1$. In such instances, the null hypothesis is that the coefficients in an AR(p) error model for e_t are equal to the ARDL coefficients on the lagged y's, extra lags are included in the test procedure, and the chi-square statistic has p degrees of freedom; it is equal to the number of observations used to estimate the test equation multiplied by that equation's R^2.

EXAMPLE 9.16 | A Consumption Function

Suppose that consumption expenditure C is a linear function of "permanent" income Y^*

$$C_t = \omega + \beta Y_t^*$$

Permanent income is unobserved. We will assume that it consists of a trend term and a geometrically weighted average of observed current and past incomes, $Y_t, Y_{t-1}, \ldots$

$$Y_t^* = \gamma_0 + \gamma_1 t + \gamma_2\left(Y_t + \lambda Y_{t-1} + \lambda^2 Y_{t-2} + \lambda^3 Y_{t-3} + \cdots\right)$$

where $t = 0, 1, 2, \ldots$ is the trend term. In this model, consumers anticipate that their income will trend, presumably upwards, adjusted by a weighted average of their past incomes. For reasons that will become apparent in Chapter 12, it is convenient to consider a differenced version of the model where we relate the change in consumption $DC_t = C_t - C_{t-1}$ to the change in actual income $DY_t = Y_t - Y_{t-1}$. This version of the model can be written as

$$\begin{aligned} DC_t = C_t - C_{t-1} &= \left(\omega + \beta Y_t^*\right) - \left(\omega + \beta Y_{t-1}^*\right) = \beta\left(Y_t^* - Y_{t-1}^*\right) \\ &= \beta\Big\{\gamma_0 + \gamma_1 t + \gamma_2\left(Y_t + \lambda Y_{t-1} + \lambda^2 Y_{t-2} + \lambda^3 Y_{t-3} + \cdots\right) \\ &\quad - \Big[\gamma_0 + \gamma_1(t-1) + \gamma_2\left(Y_{t-1} + \lambda Y_{t-2} + \lambda^2 Y_{t-3}\right. \\ &\quad \left. + \lambda^3 Y_{t-4} + \cdots\right)\Big]\Big\} \\ &= \beta\gamma_1 + \beta\gamma_2\left(DY_t + \lambda DY_{t-1} + \lambda^2 DY_{t-2} + \lambda^3 DY_{t-3} + \cdots\right) \end{aligned}$$

Setting $\alpha = \beta\gamma_1$ and $\beta_0 = \beta\gamma_2$ and adding an error term, this equation, in more familiar notation, becomes

$$DC_t = \alpha + \beta_0\left(DY_t + \lambda DY_{t-1} + \lambda^2 DY_{t-2} + \lambda^3 DY_{t-3} + \cdots\right) + e_t \tag{9.75}$$

Its ARDL(1, 0) representation is

$$DC_t = \delta + \lambda DC_{t-1} + \beta_0 DY_t + v_t \tag{9.76}$$

To estimate this model, we use quarterly data on Australian consumption expenditure and national disposable income from 1959Q3 to 2016Q3, stored in the data file *cons_inc*. Estimating (9.76) yields

$$\begin{aligned} \widehat{DC}_t &= 478.6 + 0.3369 DC_{t-1} + 0.0991 DY_t \\ \text{(se)} \quad & (74.2) \quad (0.0599) \qquad (0.0215) \end{aligned}$$

The delay multipliers from this model are 0.0991, 0.0334, 0.0112, …. The total multiplier is $0.0991/(1 - 0.3369) = 0.149$. At first, these values may seem low for what could be interpreted as a marginal propensity to consume. However, because a trend term is included in the model, we are measuring departures from that trend. The LM test for serial correlation in the errors described in Section 9.4.2 was conducted for lags 1, 2, 3, and 4; in each case, a null hypothesis of no serial correlation was not rejected at a 5% significance level. To see if this lack of serial correlation in the errors could be attributable to an AR(1) model with parameter λ for the errors in (9.75), the steps for the test in the previous subsection were followed, yielding a test value of $\chi^2 = (T-1) \times R^2 = 227 \times 0.00025 = 0.057$. Given the 5% significance level for a $\chi^2_{(1)}$-distribution is 3.84, we fail to reject the null hypothesis that the errors in the IDL representation can be described by the process $e_t = \lambda e_{t-1} + v_t$. Put another way, there is no evidence to suggest that the existence of an MA(1) error of the form $v_t = e_t - \lambda e_{t-1}$ is a source of inconsistency in the estimation of (9.76).

Deriving Multipliers from an ARDL Representation

The geometrically declining lag model is a convenient one if we believe the lag weights do in fact satisfy, or approximately satisfy, the restrictions $\beta_s = \lambda^s \beta_0$. However, there are many other lag patterns that may be realistic. The largest impact of a change in an explanatory variable may not be felt immediately; the lag weights may increase at first and then decline. How do we decide what might be reasonable restrictions to impose? Instead of beginning with the IDL representation and choosing restrictions a priori, an alternative strategy is to begin with an ARDL representation whose lags have been chosen using conventional model selection criteria and to derive the restrictions on the IDL model implied by the chosen ARDL model. Specifically, we first estimate the finite number of θ's and δ's from an ARDL model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + v_t \tag{9.77}$$

For these estimates to be compatible with the infinite number of β's in the IDL model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \cdots + e_t \tag{9.78}$$

restrictions have to be placed on the β's. The strategy is to find expressions for the β's in terms of the θ's and δ's such that equations (9.77) and (9.78) are equivalent. One way to do so is to use recursive substitution, substituting out the lagged dependent variables on the right-hand side of (9.77), and going back indefinitely. This process becomes messy very quickly, however, particularly when there are several lags. Our task for the general case is made much easier if we can master some heavy machinery known as the **lag operator**.

The lag operator L has the effect of lagging a variable,

$$Ly_t = y_{t-1}$$

For lagging a variable twice, we have

$$L(Ly_t) = Ly_{t-1} = y_{t-2}$$

which we write as $L^2 y_t = y_{t-2}$. More generally, L raised to the power of s means lag a variable s times

$$L^s y_t = y_{t-s}$$

Now we are in a position to write the ARDL model in terms of lag operator notation. Equation (9.77) becomes

$$y_t = \delta + \theta_1 L y_t + \theta_2 L^2 y_t + \cdots + \theta_p L^p y_t + \delta_0 x_t + \delta_1 L x_t + \delta_2 L^2 x_t + \cdots + \delta_q L^q x_t + v_t \tag{9.79}$$

Bringing the terms that contain y_t to the left side of the equation and factoring out y_t and x_t yields

$$\left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_p L^p\right) y_t = \delta + \left(\delta_0 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_q L^q\right) x_t + v_t \tag{9.80}$$

This algebra is starting to get heavy. It will be easier if we continue in terms of a specific example.

EXAMPLE 9.17 | Deriving Multipliers for an Infinite Lag Okun's Law Model

In Example 9.13, using data from the file *okun5_aus*, we estimated a finite distributed lag model for Okun's Law, with the change in unemployment DU_t related to the current value and four lags of GDP growth, $G_t, G_{t-1}, \ldots, G_{t-4}$. Suppose, instead, that we wanted to entertain an IDL with values for G going back into the indefinite past. The estimates in Table 9.7 suggest a geometrically declining lag would be inappropriate. The estimated coefficient for G_{t-1} is larger (in absolute value) than that for G_t and then the coefficients decline. To decide on what might be a suitable lag distribution, we begin by estimating an ARDL model. After experimenting with different values for p and q, taking into consideration significance of the coefficient estimates and the possibility of serial correlation in the errors, we settled on the ARDL(2, 1) model

$$DU_t = \delta + \theta_1 DU_{t-1} + \theta_2 DU_{t-2} + \delta_0 G_t + \delta_1 G_{t-1} + v_t \tag{9.81}$$

Using the lag operator notation in (9.80), this equation can be written as

$$\left(1 - \theta_1 L - \theta_2 L^2\right) DU_t = \delta + \left(\delta_0 + \delta_1 L\right) G_t + v_t \tag{9.82}$$

Now suppose that it is possible to define an inverse of $\left(1 - \theta_1 L - \theta_2 L^2\right)$, that we write as $\left(1 - \theta_1 L - \theta_2 L^2\right)^{-1}$, which is such that

$$\left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} \left(1 - \theta_1 L - \theta_2 L^2\right) = 1$$

This concept is a bit abstract, but we do not have to figure the inverse out. Using it will seem like magic the first time that you encounter it. Stick with us. We have nearly reached the essential result. Multiplying both sides of (9.82) by $\left(1 - \theta_1 L - \theta_2 L^2\right)^{-1}$ yields

$$\begin{aligned} DU_t = &\left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} \delta \\ &+ \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} \times \left(\delta_0 + \delta_1 L\right) G_t \\ &+ \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} v_t \end{aligned} \tag{9.83}$$

This representation is useful because we can equate it with the IDL representation

$$\begin{aligned} DU_t &= \alpha + \beta_0 G_t + \beta_1 G_{t-1} + \beta_2 G_{t-2} + \beta_3 G_{t-3} + \cdots + e_t \\ &= \alpha + \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots\right) G_t + e_t \end{aligned} \tag{9.84}$$

For (9.83) and (9.84) to be identical, it must be true that

$$\alpha = \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} \delta \tag{9.85}$$

$$\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots = \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} \left(\delta_0 + \delta_1 L\right) \tag{9.86}$$

$$e_t = \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} v_t \tag{9.87}$$

Equation (9.85) can be used to derive α in terms of θ_1, θ_2, and δ, and equation (9.86) can be used to derive the β's in terms of the θ's and δ's. To see how, first multiply both sides of (9.85) by $\left(1 - \theta_1 L - \theta_2 L^2\right)$ to obtain $\left(1 - \theta_1 L - \theta_2 L^2\right)\alpha = \delta$. Then, recognizing that the lag of a constant is the same constant ($L\alpha = \alpha$), we have

$$\left(1 - \theta_1 - \theta_2\right)\alpha = \delta \quad \text{and} \quad \alpha = \frac{\delta}{1 - \theta_1 - \theta_2}$$

Turning now to the β's, we multiply both sides of (9.86) by $\left(1 - \theta_1 L - \theta_2 L^2\right)$ to obtain

$$\begin{aligned}\delta_0 + \delta_1 L &= \left(1 - \theta_1 L - \theta_2 L^2\right)\left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots\right) \\ &= \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots \\ &\quad -\theta_1\beta_0 L - \theta_1\beta_1 L^2 - \theta_1\beta_2 L^3 - \cdots \\ &\quad - \theta_2\beta_0 L^2 - \theta_2\beta_1 L^3 - \cdots \\ &= \beta_0 + \left(\beta_1 - \theta_1\beta_0\right)L + \left(\beta_2 - \theta_1\beta_1 - \theta_2\beta_0\right)L^2 \\ &\quad + \left(\beta_3 - \theta_1\beta_2 - \theta_2\beta_1\right)L^3 + \cdots\end{aligned} \tag{9.88}$$

Notice how we can do algebra with the lag operator. We have used the fact that $L^r L^s = L^{r+s}$.

Equation (9.88) holds the key to deriving the β's in terms of the θ's and the δ's. For both sides of this equation to mean the same thing (to imply the same lags), coefficients of like powers in the lag operator must be equal. To make what follows more transparent, we rewrite (9.88) as

$$\begin{aligned}&\delta_0 + \delta_1 L + 0L^2 + 0L^3 \\ &\quad = \beta_0 + \left(\beta_1 - \theta_1\beta_0\right)L + \left(\beta_2 - \theta_1\beta_1 - \theta_2\beta_0\right)L^2 \\ &\qquad + \left(\beta_3 - \theta_1\beta_2 - \theta_2\beta_1\right)L^3 + \cdots\end{aligned} \tag{9.89}$$

Equating coefficients of like powers in L yields

$$\begin{aligned}\delta_0 &= \beta_0 \\ \delta_1 &= \beta_1 - \theta_1\beta_0 \\ 0 &= \beta_2 - \theta_1\beta_1 - \theta_2\beta_0 \\ 0 &= \beta_3 - \theta_1\beta_2 - \theta_2\beta_1\end{aligned}$$

and so on. Thus, the β's can be found from the θ's and the δ's using the recursive equations

$$\begin{aligned}\beta_0 &= \delta_0 \\ \beta_1 &= \delta_1 + \theta_1\beta_0 \\ \beta_j &= \theta_1\beta_{j-1} + \theta_2\beta_{j-2} \quad \text{for} \quad j \geq 2\end{aligned} \tag{9.90}$$

You are probably asking: Do I have to go through all this each time I want to derive some multipliers for an ARDL model? The answer is no. You can start from the equivalent of equation (9.88) which, in its general form, is

$$\begin{aligned}\delta_0 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_q L^q = &\left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_p L^p\right) \\ &\times \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots\right)\end{aligned} \tag{9.91}$$

Given the values p and q for your ARDL model, you need to multiply out the above expression, and then equate coefficients of like powers in the lag operator.

EXAMPLE 9.18 | Computing the Multiplier Estimates for the Infinite Lag Okun's Law Model

Using the data file *okun5_aus*, the estimated ARDL(2,1) model for Okun's Law is

$$\begin{aligned}\widehat{DU_t} &= 0.1708 + 0.2639 DU_{t-1} + 0.2072 DU_{t-2} \\ (\text{se}) \quad &(0.0328) \quad (0.0767) \qquad (0.0720) \\ &- 0.0904 G_t - 0.1296 G_{t-1} \\ &(0.0244) \quad (0.0252)\end{aligned} \tag{9.92}$$

Using the relationships in (9.90), the impact multiplier and the delay multipliers for the first 4 quarters are given by[16]

$$\begin{aligned}\hat{\beta}_0 &= \hat{\delta}_0 = -0.0904 \\ \hat{\beta}_1 &= \hat{\delta}_1 + \hat{\theta}_1\hat{\beta}_0 = -0.129647 - 0.263947 \times 0.090400 \\ &= -0.1535 \\ \hat{\beta}_2 &= \hat{\theta}_1\hat{\beta}_1 + \hat{\theta}_2\hat{\beta}_0 = -0.263947 \times 0.153508 \\ &\quad - 0.207237 \times 0.090400 = -0.0593\end{aligned}$$

[16] In the calculations, we carry the values to six decimal places to minimize rounding error.

$$\hat{\beta}_3 = \hat{\theta}_1\hat{\beta}_2 + \hat{\theta}_2\hat{\beta}_1 = -0.263947 \times 0.059252 - 0.207237 \times 0.153508 = -0.0475$$

$$\hat{\beta}_4 = \hat{\theta}_1\hat{\beta}_3 + \hat{\theta}_2\hat{\beta}_2 = -0.263947 \times 0.047452 - 0.207237 \times 0.059252 = -0.0248$$

An increase in GDP growth leads to a fall in unemployment. The effect increases from the current quarter to the next quarter, declines dramatically after that and then gradually declines to zero. This property—that the weights at long lags go to zero—is an essential one for the above analysis to be valid. The weights are displayed in Figure 9.12 for lags up to 10 quarters.

To estimate the total multiplier that is given by $\sum_{j=0}^{\infty}\beta_j$, we can sum the progressions implied by (9.90), but an easier way is to assume the process is in long-run equilibrium with no changes in DU and G, and to examine the effect of a change in G on the long-run equilibrium. Being in log-run equilibrium means we can ignore the time subscript and the error term in (9.92), giving

$$DU = 0.1708 + 0.2639DU + 0.2072DU - 0.0904G - 0.1296G$$

or

$$DU = \frac{0.1708 - (0.0904 + 0.1296)G}{1 - 0.2639 - 0.2072} = 0.3229 - 0.4160G$$

The total multiplier is given by $d(DU)/dG = -0.416$. The sum of the lag coefficients in Figure 9.12 is $\sum_{s=0}^{10}\hat{\beta}_s = -0.414$; most of the impact of a change in G is felt in the first 10 quarters. An estimate of the normal growth rate that is needed to maintain a constant rate of unemployment is $\hat{G}_N = -\hat{\alpha}/\sum_{j=0}^{\infty}\hat{\beta}_j = 0.3229/0.416 = 0.78\%$. The total multiplier estimate from the finite distributed lag model was higher in absolute value at -0.528, but the estimate of the normal growth rate was the same at 0.78%.

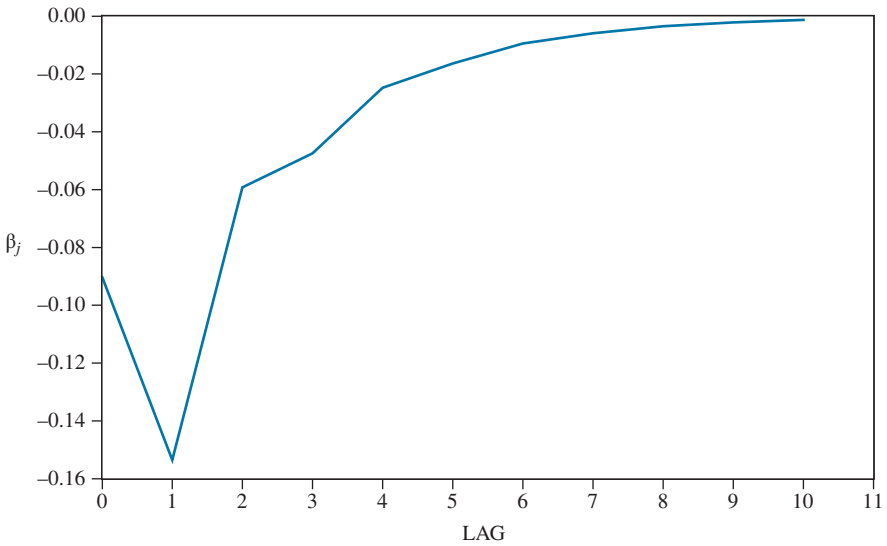


FIGURE 9.12 **Lag distribution from Okun's Law ARDL(2, 1) model.**

The Error Term

In Example 9.18, we used least squares to estimate the ARDL model and conveniently ignored the error term. The question we need to ask is whether the error term will be such that the least squares estimator is consistent. In equation (8.47), we found that

$$e_t = \left(1 - \theta_1 L - \theta_2 L^2\right)^{-1} v_t$$

Multiplying both sides of this equation by $\left(1 - \theta_1 L - \theta_2 L^2\right)$ gives

$$\left(1 - \theta_1 L - \theta_2 L^2\right) e_t = v_t$$

$$e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} = v_t$$

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + v_t$$

In the general ARDL(p, q) model, this equation becomes

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} + v_t \tag{9.93}$$

For v_t to be uncorrelated, which is required for least squares estimation of the ARDL model to be consistent, the errors e_t must satisfy (9.93). That is, they must follow an AR(p) process with the same coefficients as in the AR component of the ARDL model. The test for consistency of least squares described earlier in the context of the geometric lag model can be extended to the general case.

EXAMPLE 9.19 | Testing for Consistency of Least Squares Estimation of Okun's Law

The starting point for this test is the assumption that the errors e_t in the IDL representation follow an AR(2) process

$$e_t = \psi_1 e_{t-1} + \psi_2 e_{t-2} + v_t$$

with the v_t being uncorrelated. Then, given the ARDL representation

$$DU_t = \delta + \theta_1 DU_{t-1} + \theta_2 DU_{t-2} + \delta_0 G_t + \delta_1 G_{t-1} + v_t \tag{9.94}$$

the null hypothesis is $H_0: \psi_1 = \theta_1,\ \psi_2 = \theta_2$. To find the test statistic, we compute $\hat{e}_t = \hat{\theta}_1 \hat{e}_{t-1} + \hat{\theta}_2 \hat{e}_{t-2} + \hat{u}_t$ where the $\hat{u}_t$ are the residuals from the estimated equation in (9.92). Then, regressing $\hat{u}_t$ on a constant, DU_{t-1}, DU_{t-2}, G_t, G_{t-1}, $\hat{e}_{t-1}$, and $\hat{e}_{t-2}$ yields $R^2 = 0.02089$ and a test value $\chi^2 = (T-3) \times R^2 = 150 \times 0.02089 = 3.13$. The 5% critical value is $\chi^2_{(0.95,2)} = 5.99$ implying we fail to reject H_0 at a 5% significance level. There is not sufficient evidence to conclude that serially correlated errors are a source of inconsistency in least squares estimation of (9.94).

Assumptions for the Infinite Distributed Lag Model Several assumptions underlie least squares estimation of the consumption function and Okun's Law examples. Here we summarize those assumptions and discuss implications of variations of them.

IDL1: The time series y and x are stationary and weakly dependent.

IDL2: The infinite distributed lag model describing how y responds to current and past values of x can be written as

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + e_t \tag{9.95}$$

with $\beta_s \to 0$ as $s \to \infty$.

IDL3: Corresponding to (9.95) is an ARDL(p, q) model

$$y_t = \delta + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + v_t \tag{9.96}$$

where $v_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_p e_{t-p}$.

IDL4: The errors e_t are strictly exogenous,

$$E(e_t|\mathbf{X}) = 0$$

where $\mathbf{X}$ includes all current, past, and future values of x.

IDL5: The errors e_t follow the AR(p) process

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} + u_t$$

where

i. u_t is exogenous with respect to current and past values of x and past values of y,

$$E(u_t|x_t, x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \ldots) = 0$$

ii. u_t is homoskedastic, $\text{var}(u_t|x_t) = \sigma_u^2$

Under assumptions IDL2 and IDL3, expressions for the lag weights β_s in terms of the parameters θ's and δ's can be found by equating coefficients of like powers of the lag operator in the product

$$\delta_0 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_q L^q = \left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_p L^p\right) \times \left(\beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \cdots\right) \tag{9.97}$$

The assumption IDL5 is a very special case of an autocorrelated error model for (9.95) and for that reason we described a test of its validity. It is required for least squares estimation of (9.96) to be consistent. Because the exogeneity assumption IDL5(i) includes all past values of y, it is sufficient to ensure v_t will not be autocorrelated; IDL5(ii) is needed for OLS standard errors to be valid. If IDL5 holds and least squares estimates of (9.96) are used to find estimates of the β's through equation (9.97), strict exogeneity for e_t (IDL4) is required for the β's to have a causal interpretation. This requirement is similar to that for nonlinear least squares and generalized least squares estimation of the autocorrelated error model.

An alternative assumption to IDL5 is

IDL5*: The errors e_t are uncorrelated, $\text{cov}\left(e_t, e_s | x_t, x_s\right) = 0$ for $t \neq s$ and homoskedastic, $\text{var}\left(e_t | x_t\right) = \sigma_e^2$.

In this case, the errors $v_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_p e_{t-p}$ follow an MA(p) process, and least squares estimation of (9.96) is inconsistent. The instrumental variables approach studied in Chapter 10 can be used as an alternative.

Finally, we note that both an FDL model with autocorrelated errors and an IDL model can be transformed to ARDL models. Thus, an issue that arises after estimating an ARDL model is whether to interpret it as an FDL model with autocorrelated errors or an IDL model. An attractive way out of this dilemma is to assume an FDL model and use HAC standard errors. In many cases, an IDL model will be well approximated by an FDL, and using HAC standard errors avoids having to make the restrictive strict exogeneity assumption.

9.6 Exercises

9.6.1 Problems

9.1 **a.** Show that the mean-squared forecast error $E\left[\left(\hat{y}_{T+1} - y_{T+1}\right)^2 \middle| I_T\right]$ for a forecast $\hat{y}_{T+1}$, that depends only on past information I_T, can be written as

$$E\left[\left(\hat{y}_{T+1} - y_{T+1}\right)^2 \middle| I_T\right] = E\left[\left\{\left(\hat{y}_{T+1} - E(y_{T+1}|I_T)\right) - \left(y_{T+1} - E(y_{T+1}|I_T)\right)\right\}^2 \middle| I_T\right]$$

b. Show that $E\left[\left(\hat{y}_{T+1} - y_{T+1}\right)^2 \middle| I_T\right]$ is minimized by choosing $\hat{y}_{T+1} = E\left(y_{T+1}|I_T\right)$.

9.2 Consider the AR(1) model $y_t = \delta + \theta y_{t-1} + e_t$ where $|\theta| < 1, E\left(e_t|I_{t-1}\right) = 0$ and $\text{var}\left(e_t|I_{t-1}\right) = \sigma^2$. Let $\bar{y}_{-1} = \sum_{t=2}^{T} y_t/(T-1)$ (the average of the observations on y with the first one missing) and $\bar{y}_{-T} = \sum_{t=2}^{T} y_{t-1}/(T-1)$ (the average of the observations on y with the last one missing).

a. Show that the least squares estimator for θ can be written as

$$\hat{\theta} = \theta + \frac{\sum_{t=2}^{T} e_t\left(y_{t-1} - \bar{y}_{-T}\right)}{\sum_{t=2}^{T}\left(y_{t-1} - \bar{y}_{-T}\right)^2}$$

b. Explain why $\hat{\theta}$ is a biased estimator for θ.

c. Explain why $\hat{\theta}$ is a consistent estimator for θ.

9.3 Consider a stationary model that combines the AR(2) model $y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + e_t$ with an AR(1) error model $e_t = \rho e_{t-1} + v_t$ where $E(v_t|I_{t-1}) = 0$. Show that

$$E(y_t|I_{t-1}) = \delta(1-\rho) + (\theta_1 + \rho)\, y_{t-1} + (\theta_2 - \theta_1\rho)y_{t-2} - \theta_2\rho y_{t-3}$$

Why will the assumption $E(y_t|I_{t-1}) = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2}$ be violated if the errors are autocorrelated?

9.4 Consider the ARDL(2, 1) model

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \delta_1 x_{t-1} + e_t$$

with auxiliary AR(1) model $x_t = \alpha + \phi x_{t-1} + v_t$, where $I_t = \{y_t, y_{t-1}, \ldots, x_t, x_{t-1}, \ldots\}$, $E(e_t|I_{t-1}) = 0$, $E(v_t|I_{t-1}) = 0$, $\text{var}(e_t|I_{t-1}) = \sigma_e^2$, $\text{var}(v_t|I_{t-1}) = \sigma_v^2$, and v_t and e_t are independent. Assume that sample observations are available for $t = 1, 2, \ldots, T$.

a. Show that the best forecasts for periods $T+1$, $T+2$ and $T+3$ are given by

$$\hat{y}_{T+1} = \delta + \theta_1 y_T + \theta_2 y_{T-1} + \delta_1 x_T$$
$$\hat{y}_{T+2} = \delta + \delta_1\alpha + \theta_1\hat{y}_{T+1} + \theta_2 y_T + \delta_1\phi x_T$$
$$\hat{y}_{T+3} = \delta + \delta_1\alpha + \delta_1\phi\alpha + \theta_1\hat{y}_{T+2} + \theta_2\hat{y}_{T+1} + \delta_1\phi^2 x_T$$

b. Show that the variances of the forecast errors are given by

$$\sigma_{f1}^2 = E\left(\left(y_{T+1} - \hat{y}_{T+1}\right)^2 \middle| I_T\right) = \sigma_e^2$$
$$\sigma_{f2}^2 = E\left(\left(y_{T+2} - \hat{y}_{T+2}\right)^2 \middle| I_T\right) = \left(1 + \theta_1^2\right)\sigma_e^2 + \delta_1^2\sigma_v^2$$
$$\sigma_{f3}^2 = E\left(\left(y_{T+3} - \hat{y}_{T+3}\right)^2 \middle| I_T\right) = \left(\left(\theta_1^2 + \theta_2\right)^2 + \theta_1^2 + 1\right)\sigma_e^2 + \delta_1^2\left(\left(\theta_1 + \phi\right)^2 + 1\right)\sigma_v^2$$

9.5 Let e_t denote the error term in a time series regression. We wish to compare the autocorrelations from an AR(1) error model $e_t = \rho e_{t-1} + v_t$ with those from an MA(1) error model $e_t = \phi v_{t-1} + v_t$. In both cases, we assume that $E(v_t v_{t-s}) = 0$ for $s \neq 0$ and $E(v_t^2) = \sigma_v^2$. Let $\rho_s = E(e_t e_{t-s})/\text{var}(e_t)$ be the s-th order autocorrelation for e_t. Show that,

a. for an AR(1) error model, $\rho_1 = \rho, \rho_2 = \rho^2, \rho_3 = \rho^3, \ldots$
b. for an MA(1) error model, $\rho_1 = \phi/\left(1 + \phi^2\right), \rho_2 = 0, \rho_3 = 0, \ldots$

Describe in words the difference between the two autocorrelation structures.

9.6 This question is designed to clarify some of the results used to explain HAC standard errors.

a. Given that $\widehat{\text{var}}(\hat{q}_t) = (T-2)^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^2\hat{e}_t^2$ and $s_x^2 = T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^2$, show that

$$\frac{\sum_{t=1}^{T}\widehat{\text{var}}(\hat{q}_t)}{T^2\left(s_x^2\right)^2} = \frac{T\sum_{t=1}^{T}(x_t - \bar{x})^2\hat{e}_t^2}{(T-2)\left(\sum_{t=1}^{T}(x_t - \bar{x})^2\right)^2}$$

b. For $T = 4$, write out all the terms in the summations

$$\text{(i)}\ \sum_{t=1}^{T-1}\sum_{s=1}^{T-t}\text{cov}(q_t, q_{t+s}) \quad \text{and} \quad \text{(ii)}\ \sum_{s=1}^{T-1}(T-s)\,\text{cov}(q_t, q_{t+s})$$

What assumption is necessary for these two summations to be equal?

c. For the simple regression model $y_t = \alpha + \beta_0 x_t + e_t$ with $E(e_t|x_t) = 0$ show that $\text{cov}(e_t, e_s|x_t, x_s) = 0$ for $t \neq s$ implies $\text{cov}(q_t, q_s) = 0$ where $q_t = (x_t - \mu_x)\, e_t$.

9.7 In Section 9.5.3, we described how a generalized least squares (GLS) estimator for α and β_0 in the regression model $y_t = \alpha + \beta_0 x_t + e_t$, with AR(1) errors $e_t = \rho e_{t-1} + v_t$ and known ρ, can be computed by applying OLS to the transformed model $y_t^* = \alpha^* + \beta_0 x_t^* + v_t$ where $y_t^* = y_t - \rho y_{t-1}$, $\alpha^* = \alpha(1-\rho)$ and $x_t^* = x_t - \rho x_{t-1}$. In large samples, the GLS estimator is minimum variance because the v_t are homoskedastic and not autocorrelated. However, x_t^* and y_t^* can only be found for $t = 2, 3, \ldots, T$. One observation is lost through the transformation. To ensure the GLS estimator is minimum variance in small samples, a transformed observation for $t = 1$ has to be included. Let $e_1^* = \sqrt{1-\rho^2}e_1$.

a. Using results in Appendix 9B, show that $\text{var}(e_1^*) = \sigma_v^2$ and that e_1^* is uncorrelated with v_t, $t = 2, 3, \ldots, T$.

b. Explain why the result in (a) implies OLS applied to the following transformed model will yield a minimum variance estimator

$$y_t^* = \alpha j_t + \beta_0 x_t^* + e_t^*$$

where $y_t^* = y_t - \rho y_{t-1}, j_t = 1 - \rho, x_t^* = x_t - \rho x_{t-1}$, and $e_t^* = e_t - \rho e_{t-1} = v_t$ for $t = 2, 3, \ldots, T$, and, for $t = 1$, $y_1^* = \sqrt{1-\rho^2}y_1, j_1 = \sqrt{1-\rho^2}$, and $x_1^* = \sqrt{1-\rho^2}x_1$. This estimator, particularly when it is used iteratively with an estimate of ρ, is often known as the Prais–Winsten estimator.

9.8 Consider the following distributed lag model relating the percentage growth in private investment (*INVGWTH*) to the federal funds rate of interest (*FFRATE*).

$$INVGWTH_t = 4 - 0.4FFRATE_t - 0.6FFRATE_{t-1} - 0.3FFRATE_{t-2} - 0.2FFRATE_{t-3}$$

a. Suppose $FFRATE = 1\%$ for $t = 1, 2, 3, 4$. Use the abovementioned equation to forecast *INVGWTH* for $t = 4$.

b. Suppose that *FFRATE* is raised by one percentage point to 2% in period $t = 5$ and then returned to its original level of 1% for $t = 6, 7, 8, 9$. Use the equation to forecast *INVGWTH* for periods $t = 5, 6, 7, 8, 9$. Relate the changes in your forecasts to the values of the coefficients. What are the delay multipliers?

c. Suppose that *FFRATE* is raised to 2% for periods $t = 5, 6, 7, 8, 9$. Use the equation to forecast *INVGWTH* for periods $t = 5, 6, 7, 8, 9$. Relate the changes in your forecasts to the values of the coefficients. What are the interim multipliers? What is the total multiplier?

9.9 Using 157 weekly observations on sales revenue (*SALES*) and advertising expenditure (*ADV*) in millions of dollars for a large department store, the following relationship was estimated

$$\widehat{SALES}_t = 18.74 + 1.006ADV_t + 3.926ADV_{t-1} + 2.372ADV_{t-2}$$

a. How many degrees of freedom are there for this estimated model? (Take into account the observations lost through lagged variables.)

b. Describe the relationship between sales and advertising expenditure. Include an explanation of the lagged relationship. When does advertising have its greatest impact? What is the total effect of a sustained $1 million increase in advertising expenditure?

c. The estimated covariance matrix of the coefficients is

	C	ADV_t	ADV_{t-1}	ADV_{t-2}
C	0.2927	−0.1545	−0.0511	−0.0999
ADV_t	−0.1545	0.4818	−0.3372	0.0201
ADV_{t-1}	−0.0511	−0.3372	0.7176	−0.3269
ADV_{t-2}	−0.0999	0.0201	−0.3269	0.4713

Using a two-tail test and a 5% significance level, which lag coefficients are significantly different from zero? Do your conclusions change if you use a one-tail test? Do they change if you use a 10% significance level?

d. Find 95% confidence intervals for the impact multiplier, the one-period interim multiplier, and the total multiplier.

9.10 Consider the following time series sample of size $T = 10$ on a random variable y_t whose sample mean is $\bar{y} = 0$.

t	1	2	3	4	5	6	7	8	9	10
y_t	1	4	8	5	4	−3	0	−5	−9	−5

a. Use a hand calculator or spreadsheet to compute the sample autocorrelations

$$r_1 = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_t^2} \qquad r_2 = \frac{\sum_{t=3}^{T} y_t y_{t-2}}{\sum_{t=1}^{T} y_t^2} \qquad r_3 = \frac{\sum_{t=4}^{T} y_t y_{t-3}}{\sum_{t=1}^{T} y_t^2}$$

b. Using a 5% significance level, separately test whether r_1, r_2, and r_3 are significantly different from zero. Sketch the first three bars of the correlogram. Include the significance bounds.

9.11 Using 250 quarterly observations on U.S. GDP growth (G) from 1947Q2 to 2009Q3, we calculate the following quantities.

$$\sum_{t=1}^{250}\left(G_t-\overline{G}\right)^2=333.8558 \quad \sum_{t=2}^{250}\left(G_t-\overline{G}\right)\left(G_{t-1}-\overline{G}\right)=162.9753$$

$$\sum_{t=3}^{250}\left(G_t-\overline{G}\right)\left(G_{t-2}-\overline{G}\right)=112.4882 \quad \sum_{t=4}^{250}\left(G_t-\overline{G}\right)\left(G_{t-3}-\overline{G}\right)=30.5802$$

a. Compute the first three autocorrelations $\left(r_1, r_2, \text{and } r_3\right)$ for G. Test whether each one is significantly different from zero at a 5% significance level. Sketch the first three bars of the correlogram. Include the significance bounds.

b. Given that $\sum_{t=2}^{250}\left(G_{t-1}-\overline{G}_{-1}\right)^2=333.1119$ and $\sum_{t=2}^{250}\left(G_t-\overline{G}_1\right)\left(G_{t-1}-\overline{G}_{-1}\right)=162.974$, where $\overline{G}_1=\sum_{t=2}^{250}G_t/249=1.662249$ and $\overline{G}_{-1}=\sum_{t=2}^{250}G_{t-1}/249=1.664257$, find least squares estimates of δ and θ_1 in the AR(1) model $G_t=\delta+\theta_1G_{t-1}+e_t$. Explain the difference between the estimate $\hat{\theta}_1$ and the estimate r_1 obtained in part (a).

9.12 Increases in the mortgage interest rate increase the cost of owning a house and lower the demand for houses. In this question, we use three equations to forecast the monthly change in the number of new one-family houses sold in the United States. In the first equation (XR 9.12.1), the monthly change in the number of houses *DHOMES* is regressed against two lags of the monthly change in the 30-year conventional mortgage rate *DIRATE*. In the second equation (XR 9.12.2), *DHOMES* is regressed against two lags of itself, and in the third equation (XR 9.12.3), two lags of both *DHOMES* and *DIRATE* are included as regressors.

$$DHOMES_t=\delta+\delta_1DIRATE_{t-1}+\delta_2DIRATE_{t-2}+e_{1,t} \quad \text{(XR 9.12.1)}$$

$$DHOMES_t=\delta+\theta_1DHOMES_{t-1}+\theta_2DHOMES_{t-2}+e_{2,t} \quad \text{(XR 9.12.2)}$$

$$DHOMES_t=\delta+\theta_1DHOMES_{t-1}+\theta_2DHOMES_{t-2}+\delta_1DIRATE_{t-1}+\delta_2DIRATE_{t-2}+e_{3,t} \quad \text{(XR 9.12.3)}$$

The data used are from January, 1992 (1992M1) to September, 2016 (2016M9). The units of measurement are thousands of new houses for *DHOMES* and percentage points for *DIRATE*. After differencing and allowing for two lags, three observations are lost, resulting in a total of 294 observations that were used to produce the least squares estimates in Table 9.11.

TABLE 9.11 **Coefficient Estimates for Equations for Forecasting New Houses**

	XR 9.12.1		XR 9.12.2		XR 9.12.3	
Dependent variable	$DHOMES_t$	$\hat{e}_{1,t}$	$DHOMES_t$	$\hat{e}_{2,t}$	$DHOMES_t$	$\hat{e}_{3,t}$
C	−0.92	−0.03	0.05	0.05	−1.39	0.65
$DHOMES_{t-1}$			−0.32	0.04	−0.37	0.53
$DHOMES_{t-2}$			−0.10	0.16	−0.11	0.14
$DIRATE_{t-1}$	−46.1	−0.31			−45.6	−0.003
$DIRATEt_{-2}$	−13.2	−1.17			−35.3	30.8
$\hat{e}_{1,t-1}$		−0.39		−0.05		−0.54
$\hat{e}_{1,t-2}$		−0.14		−0.17		0.03
SSE	634312	550482	599720	597568	555967	550770

a. Given $DHOMES_{2016M8}=-54$, $DHOMES_{2016M9}=18$, $DIRATE_{2016M8}=0.00$, $DIRATE_{2016M9}=0.02$, and $DIRATE_{2016M10}=-0.01$, use each of the three estimated equations to find 95% forecast intervals for $DHOMES_{2016M10}$ and $DHOMES_{2016M11}$. Comment on the results.

b. Using a 5% significance level, test for autocorrelated errors in each of the equations.

c. Using a 5% significance level, test whether *DIRATE* Granger causes *DHOMES*.

9.13 Consider the infinite lag representation $y_t = \alpha + \sum_{s=0}^{\infty} \beta_s x_{t-s} + e_t$ for the ARDL model

$$y_t = \delta + \theta_1 y_{t-1} + \theta_3 y_{t-3} + \delta_1 x_{t-1} + v_t$$

a. Show that $\alpha = \delta/(1 - \theta_1 - \theta_3)$, $\beta_0 = 0, \beta_1 = \delta_1$, $\beta_2 = \theta_1\beta_1$, $\beta_3 = \theta_1\beta_2$, and $\beta_s = \theta_1\beta_{s-1} + \theta_3\beta_{s-3}$ for $s \geq 4$.

b. Using quarterly data on U.S. inflation (*INF*), and the change in the unemployment rate (*DU*) from 1955Q2 to 2016Q1, we estimate the following version of a Phillips curve

$$\underset{\text{(se)}}{\widehat{INF}_t} = \underset{(0.049)}{0.094} + \underset{(0.051)}{0.564}INF_{t-1} + \underset{(0.052)}{0.333}INF_{t-3} - \underset{(0.084)}{0.300}DU_{t-1} \qquad SSE = 48.857$$

c. Using the results in part (a), find estimates of the first 12 lag weights in the infinite lag representation of the estimated Phillips curve in part (b). Graph those weights and comment on the graph.

d. What rate of inflation is consistent with a constant unemployment rate (where $DU = 0$ in all time periods)?

e. Let $\hat{e}_t = 0.564\hat{e}_{t-1} + 0.333\hat{e}_{t-3} + \hat{u}_t$ where the $\hat{u}_t$ are the residuals from the equation in part (b), and the initial values $\hat{e}_1, \hat{e}_2$, and $\hat{e}_3$ are set equal to zero. The *SSE* from regressing $\hat{u}_t$ on a constant, $INF_{t-1}, INF_{t-3}DU_{t-1}, \hat{e}_{t-1}$, and $\hat{e}_{t-3}$ is 47.619. Using a 5% significance level, test the hypothesis that the errors in the infinite lag representation follow the AR(3) process $e_t = \theta_1 e_{t-1} + \theta_3 e_{t-3} + v_t$. The number of observations used in this regression and that in part (b) is 241. What are the implications of this test result?

9.14 Inflationary expectations play an important role in wage negotiations between employers and employees. In this exercise, we examine how inflationary expectations of Australian businesses, collected by National Australia Bank surveys, depend on past inflation rates. The data are quarterly and run from 1989Q3 to 2016Q1. The basic model being estimated is

$$EXPN_t = \alpha + \beta_1 INF_{t-1} + e_t$$

where $EXPN_t$ is the expected percentage price increase for 3 months ahead and INF_{t-1} is the inflation rate in the previous 3 months. The left-hand panel of estimates in Table 9.12 contains OLS estimates of α and β_1 with conventional and HAC standard errors. The right-hand panel contains nonlinear least squares estimates and both sets of standard errors assuming the equation errors follow the AR(1) process $e_t = \rho e_{t-1} + v_t$. The first three sample autocorrelations of the residuals are also reported for each of the estimations.

TABLE 9.12 Estimates for Inflationary Expectations Model

	OLS Estimates			AR(1) Error Model		
		OLS	HAC		NLS	HAC
	Coefficients	Standard Errors	Standard Errors	Coefficients	Standard Errors	Standard Errors
α	1.437	0.110	0.147	1.637	0.219	0.195
β_1	0.629	0.120	0.188	0.208	0.074	0.086
ρ				0.771	0.063	0.076
		$r_1 = 0.651$			$r_1 = -0.132$	
		$r_2 = 0.466$			$r_2 = 0.099$	
		$r_3 = 0.445$			$r_3 = -0.136$	
		Observations = 106			Observations = 105	

a. What evidence is there of serial correlation in the errors e_t? What is the impact of any serial correlation on interval estimation of β_1?

b. Is there any evidence of remaining serial correlation in the errors v_t after estimating the model with an AR(1) error?

c. What is the impact of the AR(1) error assumption on the estimate for β_1? Suggest a reason for the large difference in magnitude.

d. Show that the AR(1) error model can be written as

$$EXPN_t = \delta + \theta_1 EXPN_{t-1} + \delta_1 INF_{t-1} + \delta_2 INF_{t-2} + v_t$$

where $\delta = \alpha(1-\rho), \theta_1 = \rho, \delta_1 = \beta_1$ and $\delta_2 = -\rho\beta_1$.

e. Estimating the unconstrained version of the model in part (d) via OLS yields

$$\widehat{EXPN}_t = \underset{\text{(se)}}{0.376} + \underset{(0.121)}{0.773} EXPN_{t-1} + \underset{(0.070)}{0.206} INF_{t-1} - \underset{(0.091)}{0.163} INF_{t-2} \quad (0.090)$$

Given that $\text{se}\left(\hat{\theta}_1\hat{\delta}_1 + \hat{\delta}_2\right) = 0.1045$, test the hypothesis $H_0: \theta_1\delta_1 = -\delta_2$ using a 5% significance level. What is the implication of this test result?

f. Find estimates for the first four lag coefficients of the infinite distributed lag representation of the equation estimated in part (e).

9.15 a. Write the AR(1) error model $e_t = \rho e_{t-1} + v_t$ in lag operator notation.

b. Show that

$$(1-\rho L)^{-1} = 1 + \rho L + \rho^2 L + \rho^3 L^3 + \cdots$$

and hence that

$$e_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots$$

9.6.2 Computer Exercises

9.16 Using tha data file *usmacro*, estimate the ARDL(2, 1) model

$$U_t = \delta + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \delta_1 G_{t-1} + e_t$$

Your estimates should agree with the results given in equation (9.42). Use these estimates to verify the forecast results given in Table 9.4.

9.17 Using the data file *usmacro*, estimate the AR(1) model $G_t = \alpha + \phi G_{t-1} + v_t$. From these estimates and those obtained in Exercise 9.16, use the results from Exercise 9.4 to find point and 95% interval forecasts for U_{2016Q2},U_{2016Q3}, and U_{2016Q4}.

9.18 Consider the ARDL(p, q) equation

$$U_t = \delta + \theta_1 U_{t-1} + \cdots + \theta_p U_{t-p} + \delta_1 G_{t-1} + \cdots + \delta_q G_{t-q} + e_t$$

and the data in the file *usmacro*. For $p = 2$ and $q = 1$, results from the LM test for serially correlated errors were reported in Table 9.6 for AR(k) or MA(k) alternatives with $k = 1, 2, 3, 4$. The $\chi^2 = T \times R^2$ version of the test, with missing initial values for $\hat{e}_t$ set to zero, was used to obtain those results. Considering again the model with $p = 2$ and $q = 1$, compare the results in Table 9.6 with results from the following alternative versions of the LM test.

1. The $\chi^2 = T \times R^2$ version of the test, with missing initial values for $\hat{e}_t$ dropped.
2. The F-test for the joint significance of lags of $\hat{e}_t$, with missing initial values for $\hat{e}_t$ dropped.
3. The F-test for the joint significance of lags of $\hat{e}_t$, with missing initial values for $\hat{e}_t$ set to zero.

9.19 Consider the ARDL(p, q) equation

$$U_t = \delta + \theta_1 U_{t-1} + \cdots + \theta_p U_{t-p} + \delta_1 G_{t-1} + \cdots + \delta_q G_{t-q} + e_t$$

and the data in the file *usmacro*. For $p = 2$ and $q = 1$, results from the LM test for serially correlated errors were reported in Table 9.6 for AR(k) or MA(k) alternatives with $k = 1, 2, 3, 4$. The $\chi^2 = T \times R^2$ version of the test, with missing initial values for $\hat{e}_t$ set to zero, was used to obtain those results.

a. Using the same test statistic and the same AR and MA alternatives, and a 5% significance level, test for serially correlated errors in the two models, ($p = 4, q = 3$) and ($p = 6, q = 5$).

b. Examine the residual correlograms from the two models in part (a). What do they suggest?

9.20 In Example 9.13, the following finite distributed lag model was estimated for Okun's Law using the data file *okun5_aus*.

$$DU_t = \alpha + \beta_0 G_t + \beta_1 G_{t-1} + \beta_2 G_{t-2} + \beta_3 G_{t-3} + \beta_4 G_{t-4} + e_t$$

a. Find the correlogram of the least squares residuals for this model. Is there any evidence of autocorrelation?

b. Test for autocorrelation in the residuals using the $\chi^2 = T \times R^2$ version of the LM test, with missing initial values for $\hat{e}_t$ set to zero, and lags up to 4. Is there any evidence of autocorrelation?

c. Compare 95% interval estimates for the coefficients obtained using conventional OLS standard errors with those obtained using HAC standard errors.

9.21 In Examples 9.14 and 9.15, we considered the Phillips curve

$$INF_t = INF_t^E - \gamma(U_t - U_{t-1}) + e_t = \alpha + \beta_0 DU_t + e_t$$

where inflationary expectations are assumed to be constant, $INF_t^E = \alpha$, and $\beta_0 = -\gamma$. In Example 9.15, we used data in the file *phillips5_aus* to estimate this model assuming the errors follow an AR(1) model $e_t = \rho e_{t-1} + v_t$. Nonlinear least squares estimates of the model were $\hat{\alpha} = 0.7028$, $\hat{\beta}_0 = -0.3830$, and $\hat{\rho} = 0.5001$. The equation from these estimates can be written as the following ARDL representation (see equation (9.68))

$$\begin{aligned}\widehat{INF}_t &= \hat{\alpha}(1-\hat{\rho}) + \hat{\rho} INF_{t-1} + \hat{\beta}_0 DU_t - \hat{\rho}\hat{\beta}_0 DU_{t-1} \\ &= 0.7028 \times (1 - 0.5001) + 0.5001 INF_{t-1} - 0.3830 DU_t + (0.5001 \times 0.3830)\, DU_{t-1} \\ &= 0.3513 + 0.5001 INF_{t-1} - 0.3830 DU_t + 0.1915 DU_{t-1}\end{aligned} \quad \text{(XR 9.21.1)}$$

Instead of assuming that this ARDL(1, 1) model is a consequence of an AR(1) error, another possible interpretation is that inflationary expectations depend on actual inflation in the previous quarter, $INF_t^E = \delta + \theta_1 INF_{t-1}$. If DU_{t-1} is retained because of a possible lagged effect, and we change notation so that it is line with what we are using for a general ARDL model, we have the equation

$$INF_t = \delta + \theta_1 INF_{t-1} + \delta_0 DU_t + \delta_1 DU_{t-1} + e_t \quad \text{(XR 9.21.2)}$$

a. Find least squares estimates of the coefficients in (XR 9.21.2) and compare these values with those in (XR 9.21.1). Use HAC standard errors.

b. Reestimate (XR 9.21.2) after dropping DU_{t-1}. Why is it reasonable to drop DU_{t-1}?

c. Now, suppose that inflationary expectations depend on inflation in the previous quarter and inflation in the same quarter last year, $INF_t^E = \delta + \theta_1 INF_{t-1} + \theta_4 INF_{t-4}$. Estimate the model that corresponds to this assumption.

d. Is there empirical evidence to support the model in part (c)? In your answer, consider (i) the residual correlograms from the equations estimated in parts (b) and (c), and the significance of coefficients in the complete ARDL(4, 0) model that includes INF_{t-2} and INF_{t-3}.

9.22 Using the data file *phillips5_aus*, estimate the equation

$$INF_t = \delta + \theta_1 INF_{t-1} + \theta_4 INF_{t-4} + \delta_0 DU_t + e_t$$

Assuming that the unemployment rate in 2016Q2, 2016Q3 and 2016Q4 remains constant at 6%, use the estimated equation to find 95% forecast intervals for the inflation rate in those quarters.

9.23 Using the data file *phillips5_aus*, estimate the equation

$$INF_t = \delta + \theta_1 INF_{t-1} + \theta_4 INF_{t-4} + \delta_0 DU_t + v_t$$

a. Find the first eight lag weights (delay multipliers) of the infinite distributed lag representation that corresponds to this model. What is the total multiplier?

b. Using a 5% significance level, test the hypothesis that the error term in the infinite distributed lag representation follows the AR(4) process $e_t = \theta_1 e_{t-1} + \theta_4 e_{t-4} + v_t$.

9.24 In Example 9.16, we considered a geometrically declining infinite distributed lag model to describe the relationship between the change in consumption $DC_t = C_t - C_{t-1}$ and the change in income $DY_t = Y_t - Y_{t-1}$. In this exercise, we consider instead a finite distributed lag model of the form

$$DC_t = \alpha + \sum_{s=0}^{q} \beta_s DY_{t-s} + e_t$$

a. Use the observations in the data file *cons_inc* to estimate this model assuming $q = 4$. Use HAC standard errors. Comment on (i) the distribution of the lag weights and (ii) the significance of your estimates at a 5% significance level.

b. Reestimate the equation, dropping the lags whose coefficients were not significant in part (a). Use HAC standard errors. Have there been any substantial changes in the estimates and standard errors of the coefficients of the retained lags?

c. Using an LM test with two lags, test for autocorrelation in the errors of the equation estimated in part (b). Is the use of HAC standard errors justified?

d. Assume that the errors follow the AR(1) process $e_t = \rho e_{t-1} + v_t$ with the usual assumptions on v_t. Transform the model estimated in part (b) into one that can be estimated using nonlinear least squares.

e. Use nonlinear least squares to estimate the model derived in part (d). Use HAC standard errors. Compare these estimates and their standard errors with those obtained in part (b).

f. Using the results from part (e), find an estimate for the total multiplier and its standard error. Compare these values with those obtained for the model in Example 9.16. (You will need to estimate the model in Example 9.16 to work out the standard error of its total multiplier.)

9.25 a. Using observations on the change in consumption $DC_t = C_t - C_{t-1}$ and the change in income $DY_t = Y_t - Y_{t-1}$ from 1959Q3 to 2015Q4, obtained from the data file *cons_inc*, estimate the following two models

$$DC_t = \delta + \theta_1 DC_{t-1} + \delta_0 DY_t + e_{1t}$$
$$DC_t = \alpha + \beta_0 DY_t + \beta_3 DY_{t-3} + e_{2t}$$

b. Use each model estimated in part (a) to forecast consumption C in 2016Q1, 2016Q2, and 2016Q3.

c. Use the mean-square criterion $\sum_{t=2016Q1}^{2016Q3}\left(\hat{C}_t - C_t\right)^2$ to compare the out-of-sample predictive ability of the two models.

9.26 Using time series data on five different countries, Atkinson and Leigh[17] examine changes in inequality measured as the percentage income share (*SHARE*) held by those with the top 1% of incomes. A subset of their annual data running from 1921 to 2000 can be found in the data file *inequality*.

a. It is generally recognized that inequality was high prior to the great depression, then declined during the depression and World War II, increasing again toward the end of the sample period. To capture this effect, use the observations on New Zealand to estimate the following model with a quadratic trend

$$SHARE_t = \beta_1 + \beta_2 YEAR_t + \beta_3 YEAR_t^2 + e_t$$

where $YEAR_t$ is defined as $1 = 1921, 2 = 1922, \ldots, 80 = 2000$. Plot the observations on *SHARE* and the fitted quadratic trend. Does the trend capture the general direction of the changes in *SHARE*?

b. Find the correlogram of the least-squares residuals from the equation estimated in part (a). How many of the autocorrelations (up to lag 15) are significantly different from zero at a 5% level of significance?

c. Reestimate the equation in (a) using HAC standard errors. How do they compare with the conventional standard errors? Using first the conventional coefficient covariance matrix, and then the HAC covariance matrix, find 95% interval estimates for the expected share in 2001. That is, $E(SHARE|YEAR = 81) = \beta_1 + 81\beta_2 + 81^2\beta_3$. Compare the two intervals.

d. Assuming that the errors in (a) follow the AR(1) error process $e_t = \rho e_{t-1} + v_t$, show that the model can be rewritten as [*Hint*: $YEAR_{t-1} = YEAR_t - 1$]

$$SHARE_t = \beta_1 - \rho(\beta_1 - \beta_2 + \beta_3) + \rho SHARE_{t-1} + \left[\beta_2 - \rho(\beta_2 - 2\beta_3)\right]YEAR_t + \beta_3(1-\rho)YEAR_t^2 + v_t$$

e. Estimate the equation in part (d) using nonlinear least squares. Plot the quadratic trend and compare it with that obtained in part (a).

f. Estimate the following equation using OLS and use the estimates of δ_1, δ_2, δ_3, and ρ to retrieve estimates of β_1, β_2, and β_3. How do they compare with the nonlinear least squares estimates obtained in part (e)?

$$SHARE_t = \delta_1 + \rho SHARE_{t-1} + \delta_2 YEAR_t + \delta_3 YEAR_t^2 + v_t$$

[17] Atkinson, A.B. and A. Leigh (2013), "The Distribution of Top Incomes in Five Anglo-Saxon Countries over the Long Run", *Economic Record*, 89, 1–17.

g. Find the correlogram of the least-squares residuals from the equation estimated in part (f). How many of the autocorrelations (up to lag 15) are significantly different from zero at a 5% level of significance?

h. Using the equation estimated in part (f), find a 95% interval estimate for the expected share in 2001. That is, $E(SHARE_{2001}|YEAR = 81, SHARE_{2000} = 8.25)$. Compare this interval with those obtained in part (c).

9.27 Reconsider the data file *inequality* used in Exercise 9.26 and the model in part (a) of that exercise but include the median marginal tax rate for the upper 1% of incomes (*TAX*). We are interested in whether the marginal tax rate is a useful instrument for reducing inequality. The resulting model is

$$SHARE_t = \alpha_1 + \alpha_2 TAX_t + \alpha_3 YEAR_t + \alpha_4 YEAR_t^2 + e_t$$

a. Estimate this equation using data for Canada. Obtain both conventional and HAC standard errors. Compare the 95% interval estimates for α_2 from each of the standard errors.

b. Use an LM test with a 5% significance level and three lagged residuals to test for autocorrelation in the errors of the equation estimated in part (a). What do you conclude about the use of HAC standard errors in part (a)?

c. Estimate a parameter ρ by applying OLS to the equation $\hat{e}_t = \rho\hat{e}_{t-1} + \hat{v}_t$ where $\hat{e}_t$ are the least squares residuals from part (a). What assumption is being made when you estimate this equation?

d. Transform each of the variables in the original equation using a transformation of the form $x_t^* = x_t - \hat{\rho}x_{t-1}$ and apply OLS to the transformed variables. Compute both conventional and HAC standard errors. Find the resulting 95% interval estimates for α_2. Compare them with each other and with those found in part (a).

e. Use an LM test with a 5% significance level and three lagged residuals to test for autocorrelation in the errors of the equation estimated in part (d). What do you conclude about the use of HAC standard errors in part (d)?

f. For each of the equations estimated in parts (a) and (d), discuss whether the exogeneity assumption required for consistent estimation of α_2 is likely to be satisfied.

9.28 In this exercise, we use a subset of the data compiled by Everaert and Pozzi[18] to forecast growth in per capita private consumption (*CONSN*) and growth in per capita real disposable income (*INC*) in France. Their data are annual from 1971 to 2007 and are stored in the data file *france_ep*.

a. To forecast consumption growth consider the autoregressive model

$$CONSN_t = \delta + \sum_{s=1}^{p} \theta_s CONSN_{t-s} + e_t$$

Estimate this model for $p = 1, 2, 3$, and 4. In each case, use 33 observations to ensure the same number of observations for each value of p. Based on significance of coefficients, autocorrelation in the residuals, and the Schwarz criterion, choose a suitable value for p.

b. For the choice of p in part (a), reestimate the model using all available observations and use it to find 95% interval forecasts for $CONSN_{2008}$, $CONSN_{2009}$ and $CONSN_{2010}$.

c. To forecast income growth, consider the ARDL model

$$INC_t = \delta + \sum_{s=1}^{p} \theta_s INC_{t-s} + \sum_{r=1}^{q} \delta_r HOURS_{t-r} + e_t$$

Estimate this model for $p = 1, 2$ and $q = 1, 2$. In each case, use 35 observations to ensure the same number of observations for all values of p and q. Use the Schwarz criterion to choose between the four models. In the model of your choice, are the coefficient estimates significantly different from zero at a 5% level? At a 10% level? Does the correlogram of residuals suggest that there is any serial correlation?

d. Use the model chosen in part (c) to find 95% interval forecasts for INC_{2008}, INC_{2009}, and INC_{2010}, given that $HOURS_{2008} = HOURS_{2009} = -0.0066$.

9.29 One way of modeling supply response for an agricultural crop is to specify a model in which area planted *AREA* depends on expected price, $PRICE^*$. A log-log (constant elasticity) version of this

[18] Everaert, G. and L. Ponzi (2014), "The Predictability of Aggregate Consumption Growth in OECD Countries: a Panel Data Analysis," *Journal of Applied Econometrics*, 29(3), 431–453.

model is $\ln(AREA_t) = \alpha + \gamma \ln(PRICE^*_{t+1}) + e_t$ where $PRICE^*_{t+1}$ is expected price in the next period when harvest takes place. When farmers expect price to be high, they plant more than when a low price is expected. Since they do not know the price at harvest time, we assume that they base their expectations on current and past prices, $\ln(PRICE^*_{t+1}) = \sum_{s=0}^{q} \gamma_s \ln(PRICE_{t-s})$, with more recent prices given a greater weight, $\gamma_0 > \gamma_1 > \cdots > \gamma_q$. We use this model to explain the area of sugar cane planted in a region of the Southeast Asian country of Bangladesh. Information on the delay and interim elasticities is useful for government planning. It is important to know whether existing sugar processing mills are likely to be able to handle predicted output, whether there is likely to be excess milling capacity, and whether a pricing policy linking production, processing, and consumption is desirable. Data comprising 73 annual observations on area and price are given in the data file *bangla5*.

a. Let $\beta_s = \gamma\gamma_s$. Show that the model can be written as the finite distributed lag model

$$\ln(AREA_t) = \alpha + \sum_{s=0}^{q} \beta_s \ln(PRICE_{t-s}) + e_t$$

b. Estimate the model in part (a) assuming $q = 3$. Use HAC standard errors. What are the estimated delay and interim elasticities? Comment on the results. What are the first four autocorrelations of the residuals? Are they significantly different from zero at a 5% significance level?

c. You will have discovered that the lag weights obtained in part (a) do not satisfy a priori expectations. One way to try and overcome this problem is to insist that the weights lie on a straight line

$$\beta_s = \alpha_0 + \alpha_1 s \qquad s = 0, 1, 2, 3$$

If $\alpha_0 > 0$ and $\alpha_1 < 0$, these weights will decline implying farmers place a larger weight on more recent prices when forming their expectations. Substitute $\beta_s = \alpha_0 + \alpha_{1s}$ into the original equation and hence show that this equation can be written as

$$\ln(AREA_t) = \alpha + \alpha_0 z_{t0} + \alpha_1 z_{t1} + e_t$$

where $z_{t0} = \sum_{s=0}^{3} \ln(PRICE_{t-s})$ and $z_{t1} = \sum_{s=1}^{3} s \ln(PRICE_{t-s})$.

d. Create variables z_{t0} and z_{t1} and find least squares estimates of α_0 and α_1. Use HAC standard errors.

e. Use the estimates for α_0 and α_1 to find estimates for $\beta_s = \alpha_0 + \alpha_1 s$ and comment on them. Has the original problem been cured? Do the weights now satisfy a priori expectations?

f. How do the delay and interim elasticities compare with those obtained earlier?

9.30 In this exercise, we consider a partial adjustment model as an alternative to the model used in Exercise 9.29 for modeling sugar cane area response in Bangladesh. The data are in the file *bangla5*. In the partial adjustment model long-run desired area, $AREA^*$ is a function of price,

$$AREA^*_t = \alpha + \beta_0 PRICE_t \qquad \text{(XR 9.30.1)}$$

In the short-run, fixed resource constraints prevent farmers from fully adjusting to the area desired at the prevailing price. Specifically,

$$AREA_t - AREA_{t-1} = \gamma(AREA^*_t - AREA_{t-1}) + e_t \qquad \text{(XR 9.30.2)}$$

where $AREA_t - AREA_{t-1}$ is the actual adjustment from the previous year, $AREA^*_t - AREA_{t-1}$ is the desired adjustment from the previous year, and $0 < \gamma < 1$.

a. Combine (XR 9.30.1) and (XR 9.30.2) to show that an estimable form of the model can be written as

$$AREA_t = \delta + \theta_1 AREA_{t-1} + \delta_0 PRICE_t + e_t$$

where $\delta = \alpha\gamma$, $\theta_1 = 1 - \gamma$, and $\delta_0 = \beta_0\gamma$.

b. Find least squares estimates of δ, θ_1, and δ_0. Are they significantly different from zero at a 5% significance level?

c. What are the first three autocorrelations of the residuals? Are they significantly different from zero at a 5% significance level?

d. Find estimates and standard errors for α, β_0, and γ. Are the estimates significantly different from zero at a 5% significance level?

e. Find an estimate of $AREA^*_{73}$ and compare it with $AREA_{73}$.

f. Forecast $AREA_{74}, AREA_{75}, \ldots, AREA_{80}$ assuming that price in the next 7 years does not change from the last sample value $\left(PRICE_{74} = PRICE_{75} = \cdots = PRICE_{80} = PRICE_{73}\right)$. Comment on these forecasts and compare the forecast $\widehat{AREA}_{80}$ with $AREA^*_{80}$ estimated from (XR 9.30.1).

9.31 Using data on the Maltese economy, Apap and Gravino[19] estimate a number of versions of Okun's Law. Their quarterly data run from 1999Q1 to 2012Q4 and can be found in the data file *apap*. The variables used in this exercise are $DU_t = U_t - U_{t-4}$ (the change in the unemployment rate relative to the same quarter in the previous year) and G_t (real output growth in quarter t relative to quarter $t-4$).

a. Estimate the Okun's Law equation $DU_t = \alpha + \beta_0 G_t + e_t$. Find both conventional and HAC standard errors and comment on the results.

b. Check the correlogram of the residuals $\hat{e}_t$ from the equation estimated in part (a). Is there evidence of autocorrelation?

c. Create the variable $q_t = G_t \times \hat{e}_t$, and examine its correlogram. Use this correlogram and equation (9.63) to suggest a reason why the conventional and HAC standard errors for the estimate of β_0 are similar in magnitude.

d. Estimate the finite distributed lag model

$$DU_t = \alpha + \beta_0 G_t + \beta_1 G_{t-1} + \beta_2 G_{t-2} + e_t$$

Use HAC standard errors. Is there evidence of a lagged effect of growth on unemployment? Using HAC standard errors in both cases, find a 95% interval estimate for the total multiplier and compare it with a 95% interval for the total multiplier from the model in part (a).

e. Estimate ARDL models $DU_t = \delta + \sum_{s=1}^{p} \theta_s DU_{t-s} + \sum_{r=0}^{q} \delta_r G_{t-r} + e_t$ for $p = 1, 2, 3$ and $q = 0, 1, 2$. Use HAC standard errors. Select and report the model with the largest number of lags whose coefficients are significantly different from zero at a 5% level.

f. For the model selected in part (e), find estimates for the total multiplier, the impact multiplier, and the first three delay multipliers of the infinite distributed lag representation.

g. For the model selected in part (e), find 95% interval estimates for the total multiplier and the two-period interim multiplier. How do they compare with the interval obtained in part (d)?

9.32 In their paper referred to in Exercise 9.31, Apap and Gravino examine the separate effects of output growth in the manufacturing and services sectors on changes in the unemployment rate. Their quarterly data run from 1999Q1 to 2012Q4 and can be found in the data file *apap*. The variables used in this exercise are $DU_t = U_t - U_{t-4}$ (the change in the unemployment rate relative to the same quarter in the previous year), MAN_t (real output growth in the manufacturing sector in quarter t relative to quarter $t-4$), SER_t (real output growth in the services sector in quarter t relative to quarter $t-4$), MAN_WT_t (the proportion of real output attributable to the manufacturing sector in quarter t), and SER_WT_t (the proportion of real output attributable to the services sector in quarter t). The relative effects of growth in each of the sectors on unemployment will depend not only on their growth rates but also on the relative size of each sector in the economy. To recognize this fact, construct the weighted growth variables $MAN2_t = MAN_t \times MAN\_WT_t$ and $SER2_t = SER_t \times SER\_WT_t$.

a. Use OLS with HAC standard errors to estimate the model

$$DU_t = \alpha + \gamma_0 SER2_t + \gamma_1 SER2_{t-1} + \beta_0 MAN2_t + \beta_1 MAN2_{t-1} + v_t$$

Comment on the relative importance of growth in each sector on changes in unemployment and on whether there is a lag in the effect from each sector.

b. Use an LM test with two lags and a 5% significance level to test for autocorrelation in the errors for the equation in part (a).

c. Assume that the errors in the equation in part (a) follow the AR(1) process $e_t = \rho e_{t-1} + v_t$. Show that, under this assumption, the model can be written as

$$\begin{aligned} DU_t = {} & \alpha(1-\rho) + \rho DU_{t-1} + \gamma_0 SER2_t + \left(\gamma_1 - \rho\gamma_0\right) SER2_{t-1} - \rho\gamma_1 SER2_{t-2} \\ & + \beta_0 MAN2_t + \left(\beta_1 - \rho\beta_0\right) MAN2_{t-1} - \rho\beta_1 MAN2_{t-2} + v_t \end{aligned}$$

d. Use nonlinear least squares with HAC standard errors to estimate the model in part (c). Have your conclusions made in part (a) changed?

---

[19] Apap, W. and D. Gravino (2017), "A Sectoral Approach to Okun's Law", *Applied Economics Letters* 25(5), 319–324. The authors are grateful to Wayne Apap for providing the data.

e. Use an LM test with two lags and a 5% significance level to test for autocorrelation in the errors for the equation in part (d). Is the AR(1) process adequate to model the autocorrelation in the errors of the original equation.

f. Suppose that, wanting to forecast DU_{2013Q1} using current and past information, you set up the model

$$DU_t = \delta + \theta_1 DU_{t-1} + \theta_2 DU_{t-2} + \gamma_1 SER2_{t-1} + \gamma_2 SER2_{t-2} + \delta_1 MAN2_{t-1} + \delta_2 MAN2_{t-2} + v_t$$

i. Have a sufficient number of lags of *DU* been included?

ii. Using a 5% significance level, test whether *SER2* Granger causes *DU*.

iii. Using a 5% significance level, test whether *MAN2* Granger causes *DU*.

9.33 The data file *xrate* contains monthly observations from 1986M1 to 2008M12 on the following variables[20]:

NER = the nominal exchange rate for the Australian dollar in terms of U.S. cents.
INF_AUS = the Australian inflation rate.
INF_US = the U.S. inflation rate.
DI6_AUS = the percentage change in the interest rate on an Australian government debt instrument of maturity 6 months.
DI6_US = the percentage change in the interest rate on a U.S. government debt instrument of maturity 6 months.

a. Plot *NER* against time and examine its correlogram. Does the series wander like a nonstationary series? Do the autocorrelations die out relatively quickly, suggesting a weakly dependent series?

b. Construct a variable which is the monthly change in the exchange rate, $DNER_t = NER_t - NER_{t-1}$. Plot *DNER* against time and examine its correlogram. Does the series wander like a nonstationary series? Do the autocorrelations die out relatively quickly, suggesting a weakly dependent series?

c. Theory suggests that the exchange rate will be higher when Australian inflation is low relative to that in the United States, and when the Australian interest rate is high relative to the U.S. interest rate. Construct the two variables $DINF_t = INF\_AUS_t - INF\_US_t$ and $DI6_t = DI6\_AUS_t - DI6\_US_t$, and estimate the model (using HAC standard errors)

$$DNER_t = \alpha + \beta_0 DINF_t + \beta_1 DINF_{t-1} + \gamma_0 DI6_t + \gamma_1 DI6_{t-1} + e_t$$

Comment on the results. Do the coefficients have the expected signs? Are they significantly different from zero using one-tail tests and a 5% significance level?

d. Reestimate the model in part (c), dropping variables whose coefficients had the wrong sign. Are the coefficients in the reestimated model significantly different from zero using one-tail tests and a 5% significance level? Check for serial correlation in the errors, using both the residual correlogram and an LM test with one lagged residual.

e. Reestimate the model in part (d) using feasible generalized least squares and assuming AR(1) errors. Estimate the model with both conventional and HAC standard errors. Are the coefficients in the reestimated model significantly different from zero using one-tail tests and a 5% significance level?

f. Suppose that the following model is proposed for 1-month ahead forecasting of the exchange rate

$$DNER_t = \delta + \theta_1 DNER_{t-1} + \delta_1 DINF_{t-1} + \phi_1 DI6_{t-1} + e_t$$

Estimate this model using observations from 1986M1 to 2007M12. Does it appear to be a good model for forecasting?

g. Use the model in part (f) to obtain 1-month ahead forecasts of *NER* for each of the months in 2008. (Use the actual values of $DNER_{t-1}$ to obtain each forecast.) Comment on the accuracy of the forecasts and compute the average absolute forecast error $\sum_{t=2008M1}^{2008M12} \left|\widehat{NER}_t - NER_t\right| \Big/ 12$.

9.34 In the new Keynesian Phillips curve (NKPC), inflation at time t (INF_t) depends on inflationary expectations formed at time t for time $t+1$ $(INFEX_t)$, and the output gap, defined as output less potential output. Expectations of higher inflation lead to greater inflation. The closer output is to potential

[20]These data are constructed from the data archive for Berge, T. (2014), "Forecasting Disconnected Exchange Rates," *Journal of Applied Econometrics* 29(5), 713–735.

output, the higher the inflation rate. Amberger et al.[21] compare results from estimating NKPCs with two output gaps, one that has been augmented with changes in financial variables $(FNGAP_t)$, and one that has not (GAP_t). Quarterly data for Italy for the period 1990Q1 to 2014Q4 can be found in the data file *italy*.

a. Using OLS, estimate the two equations

$$INF_t = \alpha_G + \beta_G INFEX_t + \gamma_G GAP_t + e_{Gt}$$
$$INF_t = \alpha_F + \beta_F INFEX_t + \gamma_F FNGAP_t + e_{Ft}$$

Find 95% interval estimates for γ_G and γ_F using both conventional and HAC standard errors. Comment on (i) the relative widths of the intervals with and without HAC standard errors and (ii) whether one output gap measure is preferred over another in terms of its impact on inflation.

b. What are the values of the first four residual autocorrelations from each of the two regressions in part (a)? Which ones are significantly different from zero at a 5% significance level?

c. Consider the generic equation $y_t = \alpha + \beta x_t + \gamma z_t + e_t$ with AR(2) errors $e_t = \psi_1 e_{t-1} + \psi_2 e_{t-2} + v_t$ where the v_t are not autocorrelated. Show that this model can be written as

$$y_t^* = \alpha^* + \beta x_t^* + \gamma z_t^* + v_t \qquad t = 3, 4, \ldots, T$$

where $y_t^* = y_t - \psi_1 y_{t-1} - \psi_2 y_{t-2}$, $\alpha^* = \alpha(1 - \psi_1 - \psi_2)$, $x_t^* = x_t - \psi_1 x_{t-1} - \psi_2 x_{t-2}$, and $z_t^* = z_t - \psi_1 z_{t-1} - \psi_2 z_{t-2}$.

d. Using the least squares residuals $\hat{e}_{Gt}$ from the first equation in part (a), estimate ψ_1 and ψ_2 from the regression equation $\hat{e}_{Gt} = \psi_1 \hat{e}_{G,t-1} + \psi_2 \hat{e}_{G,t-2} + \hat{v}_t$. Along the lines of the transformations in part (c), use the estimates of ψ_1 and ψ_2 to find transformed variables INF_t^*, $INFEX_t^*$, and GAP_t^* and then estimate α_G^*, β_G, and γ_G from the transformed equation $INF_t^* = \alpha_G^* + \beta_G INFEX_t^* + \gamma_G GAP_t^* + v_t$. Estimate the equation with both conventional and HAC standard errors.

e. Using the results from part (d), find 95% interval estimates for γ_G using both conventional and HAC standard errors. Comment on (i) the relative widths of the intervals with and without HAC standard errors and (ii) how the estimates and intervals compare with the corresponding ones obtained in part (a).

9.35 Do lags of the variables in the new Keynesian Phillips curve provide a good basis for forecasting quarterly inflation? In this exercise, we investigate this question using the French data from Amberger et al. See Exercise 9.34 for details. The data are stored in the data file *france*.

a. Consider ARDL models of the form

$$INF_t = \delta + \sum_{s=1}^{p} \theta_s INF_{t-s} + \sum_{r=1}^{q} \delta_r INFEX_{t-r} + \sum_{j=1}^{m} \gamma_j GAP_{t-j} + e_t$$

Using observations from 1991Q1 to 2013Q4, estimate this equation for $p = 2$, $q = 1, 2, 3, 4$ and $m = 1, 2, 3, 4$. From these 16 equations, select and report the one with the smallest value of the Schwarz criterion. Note that 92 observations should be used to estimate each equation.

b. In the equation selected in part (a), are all the estimated coefficients significantly different from zero at a 5% significance level? Does the correlogram suggest that there is no autocorrelation in the errors?

c. Use the selected model from part (a) to find 95% forecast intervals for inflation in 2014Q1, 2014Q2, 2014Q3, and 2014Q4. When computing the forecasts, use actual values of *INFEX* and *GAP* where needed but assume that the actual values of *INF* in the four forecast quarters are unknown. After you have found the forecast intervals, check whether the actual values lie within those intervals. [*Hint*: If your software does not compute standard errors of forecast errors, equation (9.41) can be used to find them for the first three quarters. For the fourth quarter, the variance of the forecast error is given by

$$\sigma_{f4}^2 = \left[\left(\theta_1^3 + 2\theta_1\theta_2\right)^2 + \left(\theta_1^2 + \theta_2\right)^2 + \theta_1^2 + 1\right]\sigma^2$$

You might like to prove this result.]

d. What assumptions are necessary for the standard errors of the forecast errors to be valid?

[21] Amberger, J., R Fendel and H. Stremmel (2017), "Improved output gaps with financial cycle information? An application to G7 countries' new Keynesian Phillips curves," *Applied Economics Letters*, 24(4), 219–228. Many thanks to Johanna Amberger for supplying the data used in this study.

9.36 Consider the following model where a dependent variable y depends on infinite distributed lags of the two variables x and z.

$$y_t = \alpha + \sum_{s=0}^{\infty} \beta_s x_{t-s} + \sum_{r=0}^{\infty} \gamma_r z_{t-r} + e_t$$

Suppose that both sets of lag weights decline geometrically, but with different parameters λ_1 and λ_2. That is, $\beta_s = \lambda_1^s \beta_0$ and $\gamma_r = \lambda_2^r \gamma_0$.

a. Show that the model can be written as

$$y_t = \alpha + \beta_0 \sum_{s=0}^{\infty} \lambda_1^s L^s x_t + \gamma_0 \sum_{r=0}^{\infty} \lambda_2^r L^r z_t + e_t$$

b. Use the result in Exercise 9.15 to show that the equation in (a) can be written as

$$\begin{aligned} y_t &= \alpha + \beta_0 (1 - \lambda_1 L)^{-1} x_t + \gamma_0 (1 - \lambda_2 L)^{-1} z_t + e_t \\ &= \alpha^* + (\lambda_1 + \lambda_2) y_{t-1} - \lambda_1 \lambda_2 y_{t-2} + \beta_0 x_t - \beta_0 \lambda_2 x_{t-1} + \gamma_0 z_t - \gamma_0 \lambda_1 z_{t-1} + v_t \end{aligned}$$

where $\alpha^* = (1 - \lambda_1)(1 - \lambda_2)\alpha$ and $v_t = e_t - (\lambda_1 + \lambda_2) e_{t-1} + \lambda_1 \lambda_2 e_{t-2}$.

c. Using data in the file *canada5*, with $y_t = INF_t$, $x_t = INFEX_t$, and $z_t = GAP_t$, estimate the last equation in part (b) using nonlinear least squares. Report the estimates, their standard errors, and one-tail p-values for a zero null hypothesis on each parameter (except the constant). Are the estimates significantly different from zero at a 5% level?

d. Find estimates of the first three lag weights for both *INFEX* and *GAP*.

e. Find estimates of the total multipliers for both *INFEX* and *GAP*.

f. Using a 5% significance level, test $H_0: \lambda_1 = \lambda_2$ versus $H_1: \lambda_1 \neq \lambda_2$. What are the implications for the model if H_0 is true?

g. The equation estimated in part (c) can be viewed as a restricted version of the more general ARDL(2, 1, 1) model

$$y_t = \alpha^* + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \delta_0 x_t + \delta_1 x_{t-1} + \phi_0 z_t + \phi_1 z_{t-1} + v_t$$

where $\frac{\delta_1}{\delta_0} \times \frac{\phi_1}{\phi_0} = -\theta_2$ and $\frac{\delta_1}{\delta_0} + \frac{\phi_1}{\phi_0} = -\theta_1$. Estimate this unrestricted model and jointly test the validity of the restrictions at a 5% level. What are the implications for the infinite distributed lags if the restrictions are not true?

h. Test the hypothesis that e_t follows an AR(2) process $e_t = (\lambda_1 + \lambda_2) e_{t-1} - \lambda_1 \lambda_2 e_{t-2} + u_t$. What are the implications of rejecting this hypothesis?

Appendix 9A The Durbin–Watson Test

In Section 9.4, two testing procedures for testing for autocorrelated errors, the sample correlogram and a Lagrange multiplier test, were considered. These are two large sample tests; their test statistics have their specified distributions in large samples. An alternative test, one that is exact in the sense that its distribution does not rely on a large sample approximation, is the Durbin–Watson test. It was developed in 1950 and for a long time was the standard test for $H_0: \rho = 0$ in the AR(1) error model $e_t = \rho e_{t-1} + v_t$. It is used less frequently today because of the need to examine upper and lower bounds, as we describe below, and because its distribution no longer holds when the equation contains a lagged dependent variable. In addition, the test is derived conditional on $\mathbf{X}$; it treats the explanatory variables as nonrandom.

It is assumed that the v_t are independent random errors with distribution $N(0, \sigma_v^2)$ and that the alternative hypothesis is one of positive autocorrelation. That is,

$$H_0: \rho = 0 \qquad H_1: \rho > 0$$

The statistic used to test H_0 against H_1 is

$$d = \frac{\sum_{t=2}^{T} \left(\hat{e}_t - \hat{e}_{t-1}\right)^2}{\sum_{t=1}^{T} \hat{e}_t^2} \tag{9A.1}$$

where the $\hat{e}_t$ are the least squares residuals $\hat{e}_t = y_t - b_1 - b_2 x_t$. To see why d is a reasonable statistic for testing for autocorrelation, we expand (9A.1) as

$$\begin{aligned} d &= \frac{\sum_{t=2}^{T} \hat{e}_t^2 + \sum_{t=2}^{T} \hat{e}_{t-1}^2 - 2\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2} \\ &= \frac{\sum_{t=2}^{T} \hat{e}_t^2}{\sum_{t=1}^{T} \hat{e}_t^2} + \frac{\sum_{t=2}^{T} \hat{e}_{t-1}^2}{\sum_{t=1}^{T} \hat{e}_t^2} - 2\frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2} \\ &\approx 1 + 1 - 2r \end{aligned} \tag{9A.2}$$

The last line in (9A.2) holds only approximately. The first two terms differ from 1 through the exclusion of $\hat{e}_1^2$ and $\hat{e}_T^2$ from the first and second numerator summations, respectively. Thus, we have

$$d \approx 2\left(1 - r_1\right) \tag{9A.3}$$

If the estimated value of ρ is $r_1 = 0$, then the Durbin–Watson statistic $d \approx 2$, which is taken as an indication that the model errors are not autocorrelated. If the estimate of ρ happened to be $r_1 = 1$ then $d \approx 0$, and thus a low value for the Durbin–Watson statistic implies that the model errors are correlated, and $\rho > 0$.

The question we need to answer is: How close to zero does the value of the test statistic have to be before we conclude that the errors are correlated? In other words, what is a critical value d_c such that we reject H_0 when $d \leq d_c$? Determination of a critical value and a rejection region for the test requires knowledge of the probability distribution of the test statistic under the assumption that the null hypothesis, $H_0 : \rho = 0$, is true. For a 5% significance level, knowledge of the probability distribution $f(d)$ under H_0 allows us to find d_c such that $P\left(d \leq d_c\right) = 0.05$. Then, as illustrated in Figure 9.A1, we reject H_0 if $d \leq d_c$ and fail to reject H_0 if $d > d_c$. Alternatively, we can state the test procedure in terms of the p-value of the test. For this one-tail test, the p-value is given by the area under $f(d)$ to the left of the calculated value of d. Thus, if the p-value is less than or equal to 0.05, it follows that $d \leq d_c$, and H_0 is rejected. If the p-value is greater than 0.05, then $d > d_c$, and H_0 is not rejected.

In any event, whether the test result is found by comparing d with d_c or by computing the p-value, the probability distribution $f(d)$ is required. A difficulty associated with $f(d)$, and one that we have not previously encountered when using other test statistics, is that this probability distribution depends on the values of the explanatory variables. Different sets of explanatory variables lead to different distributions for d. Because $f(d)$ depends on the values of the explanatory variables, the critical value d_c for any given problem will also depend on the values of the explanatory variables. This property means that it is impossible to tabulate critical values that can be used for every possible problem. With other test statistics, such as t, F, and χ^2, the tabulated critical values are relevant for all models.

There are two ways to overcome this problem. The first way is to use software that computes the p-value for the explanatory variables in the model under consideration. Instead of comparing

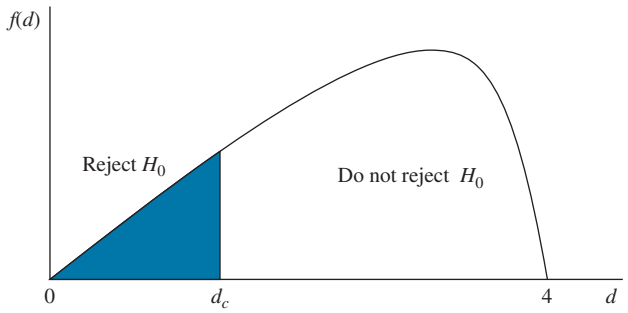


FIGURE 9.A1 **Testing for positive autocorrelation.**

the calculated d value with some tabulated values of d_c, we get our computer to calculate the p-value of the test. If this p-value is less than the specified significance level, $H_0 : \rho = 0$ is rejected, and we conclude that the errors are correlated.[1]

9A.1 The Durbin–Watson Bounds Test

In the absence of software that computes a p-value, a test known as the bounds test can be used to partially overcome the problem of not having general critical values. Durbin and Watson considered two other statistics d_L and d_U whose probability distributions do not depend on the explanatory variables and which have the property that

$$d_L < d < d_U$$

That is, irrespective of the explanatory variables in the model under consideration, d will be bounded by an upper bound d_U and a lower bound d_L. The relationship between the probability distributions $f(d_L)$, $f(d)$, and $f(d_U)$ is depicted in Figure 9.A2. Let d_{Lc} be the 5% critical value from the probability distribution for d_L. That is, d_{Lc} is such that $P(d_L \le d_{Lc}) = 0.05$. Similarly, let d_{Uc} be such that $P(d_U \le d_{Uc}) = 0.05$. Since the probability distributions $f(d_L)$ and $f(d_U)$ do not depend on the explanatory variables, it is possible to tabulate the critical values d_{Lc} and d_{Uc}. These values do depend on T and K, but it is possible to tabulate the alternative values for different T and K.

Thus, in Figure 9.A2, we have three critical values. The values d_{Lc} and d_{Uc} can be readily tabulated. The value d_c, the one in which we are really interested for testing purposes, cannot be found without a specialized computer program. However, it is clear from the figure that if the calculated value d is such that $d \le d_{Lc}$, then it must follow that $d \le d_c$, and H_0 is rejected. In addition, if $d > d_{Uc}$, then it follows that $d > d_c$, and H_0 is not rejected. If it turns out that $d_{Lc} < d < d_{Uc}$, then, because we do not know the location of d_c, we cannot be sure whether to accept or reject. These considerations led Durbin and Watson to suggest the following decision rules, known collectively as the Durbin–Watson *bounds test*:

$$\text{If } d \le d_{Lc}, \text{ reject } H_0 : \rho = 0 \text{ and accept } H_1 : \rho > 0;$$
$$\text{if } d > d_{Uc}, \text{ do not reject } H_0 : \rho = 0;$$
$$\text{if } d_{Lc} < d < d_{Uc}, \text{ the test is inconclusive.}$$

The presence of a range of values where no conclusion can be reached is an obvious disadvantage of the test. For this reason, it is preferable to have software which can calculate the required p-value if such software is available.

[1] The software packages SHAZAM and SAS, for example, will compute the exact Durbin–Watson p-value.

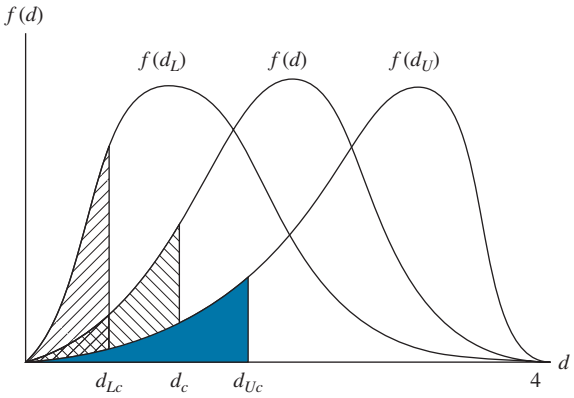


FIGURE 9.A2 **Upper and lower critical value bounds for the Durbin–Watson test.**

EXAMPLE 9.20 | Durbin–Watson Bounds Test for Phillips Curve

The 5% critical bounds for the Phillips curve in Examples 9.14 and 9.15, for $T = 117$ and $K = 2$ are[2]

$$d_{Lc} = 1.681 \qquad d_{Uc} = 1.716$$

The Durbin–Watson test value is 0.965. Since $0.965 < d_{Lc} = 1.681$, we conclude that $d < d_c$, and hence we reject $H_0: \rho = 0$; there is evidence to suggest that the errors are positively serially correlated.

Appendix 9B Properties of an AR(1) Error

We are interested in the mean, variance, and autocorrelations for e_t where $e_t = \rho e_{t-1} + v_t$ and the v_t are uncorrelated random errors with mean zero and variance σ_v^2.[3] To derive the desired properties, we begin by lagging the equation $e_t = \rho e_{t-1} + v_t$ by one period, to obtain $e_{t-1} = \rho e_{t-2} + v_{t-1}$. Then, substituting e_{t-1} into the first equation yields

$$\begin{aligned} e_t &= \rho e_{t-1} + v_t \\ &= \rho\left(\rho e_{t-2} + v_{t-1}\right) + v_t \\ &= \rho^2 e_{t-2} + \rho v_{t-1} + v_t \end{aligned} \tag{9B.1}$$

Lagging $e_t = \rho e_{t-1} + v_t$ by two periods gives $e_{t-2} = \rho e_{t-3} + v_{t-2}$. Substituting this expression for e_{t-2} into (9B.1) yields

$$\begin{aligned} e_t &= \rho^2\left(\rho e_{t-3} + v_{t-2}\right) + \rho v_{t-1} + v_t \\ &= \rho^3 e_{t-3} + \rho^2 v_{t-2} + \rho v_{t-1} + v_t \end{aligned} \tag{9B.2}$$

Repeating this process k times and rearranging the order of the lagged v's yields

$$e_t = \rho^k e_{t-k} + v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots + \rho^{k-1} v_{t-k+1} \tag{9B.3}$$

If we view the process as operating for a long time into the past, then we can let $k \to \infty$. This makes the first and last terms, $\rho^k e_{t-k}$ and $\rho^{k-1} v_{t-k+1}$, go to zero because $-1 < \rho < 1$. The result is

$$e_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots \tag{9B.4}$$

[2]These values can be found from the Durbin Watson tables on the web site principlesofeconometrics.com/poe5/poe5.htm.

[3]To simplify the exposition, we derive these properties in terms of the marginal distributions of e_t and v_t. When estimating the AR(1) error model in the body of the chapter, we make the stronger assumptions $E(v_t|\mathbf{X}) = 0$ and $\text{var}(v_t|\mathbf{X}) = \sigma_v^2$.

The regression error e_t can be written as a weighted sum of the current and past values of the uncorrelated error v_t. This is an important result. It means that all past values of the v's have an impact on the current error e_t and that this impact feeds through into y_t through the regression equation. Notice, however, that the impact of the past v's declines the further we go into the past. The weights that are attached to the lagged v's are $\rho, \rho^2, \rho^3, \ldots$. Because $-1 < \rho < 1$, these weights decline geometrically as we consider past v's that are more distant from the current period. Eventually, they become negligible.

Equation (9B.4) can be used to find the properties of the e_t. Its mean is zero, because

$$\begin{aligned} E(e_t) &= E(v_t) + \rho E(v_{t-1}) + \rho^2 E(v_{t-2}) + \rho^3 E(v_{t-3}) + \cdots \\ &= 0 + \rho \times 0 + \rho^2 \times 0 + \rho^3 \times 0 + \cdots \\ &= 0 \end{aligned}$$

To find the variance, we write

$$\begin{aligned} \operatorname{var}(e_t) &= \operatorname{var}(v_t) + \rho^2 \operatorname{var}(v_{t-1}) + \rho^4 \operatorname{var}(v_{t-2}) + \rho^6 \operatorname{var}(v_{t-3}) + \cdots \\ &= \sigma_v^2 + \rho^2 \sigma_v^2 + \rho^4 \sigma_v^2 + \rho^6 \sigma_v^2 + \cdots \\ &= \sigma_v^2 (1 + \rho^2 + \rho^4 + \rho^6 + \cdots) \\ &= \frac{\sigma_v^2}{1 - \rho^2} \end{aligned} \tag{9B.5}$$

In the abovementioned derivation, zero covariance terms are ignored because the v's are uncorrelated. The result in the last line follows from rules for the sum of a geometric progression. Using shorthand notation, we have $\sigma_e^2 = \sigma_v^2 / (1 - \rho^2)$; the variance of e depends on that for v and the value for ρ.

To find the covariance between two e's that are one period apart, we use (9B.4) and its lag to write

$$\begin{aligned} \operatorname{cov}(e_t, e_{t-1}) &= E(e_t e_{t-1}) \\ &= E\Big[(v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots) \\ &\qquad (v_{t-1} + \rho v_{t-2} + \rho^2 v_{t-3} + \rho^3 v_{t-4} \cdots) \Big] \\ &= \rho E(v_{t-1}^2) + \rho^3 E(v_{t-2}^2) + \rho^5 E(v_{t-3}^2) + \cdots \\ &= \rho \sigma_v^2 (1 + \rho^2 + \rho^4 + \cdots) \\ &= \frac{\rho \sigma_v^2}{1 - \rho^2} \end{aligned}$$

When the second line in the abovementioned derivation is expanded, only squared terms with the same subscript are retained. Because the v's are uncorrelated, the cross-product terms with different time subscripts will have zero expectation and are dropped from the third line. To obtain the fourth line from the third line, we have used $E(v_{t-k}^2) = \operatorname{var}(v_{t-k}) = \sigma_v^2$ for all lags k. In a similar way, we can show that the covariance between errors that are k periods apart is

$$\operatorname{cov}(e_t, e_{t-k}) = \frac{\rho^k \sigma_v^2}{1 - \rho^2} \qquad k > 0 \tag{9B.6}$$

From (9B.5) and (9B.6), the autocorrelations for errors that are k periods apart are given by

$$\rho_k = \operatorname{corr}(e_t, e_{t-k}) = \frac{\operatorname{cov}(e_t, e_{t-k})}{\operatorname{var}(e_t)} = \frac{\rho^k \sigma_v^2 / (1 - \rho^2)}{\sigma_v^2 / (1 - \rho^2)} = \rho^k$$

CHAPTER 10

Endogenous Regressors and Moment-Based Estimation

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Give an intuitive explanation of why correlation between a random x and the error term causes the least squares estimator to be inconsistent.
2. Describe the "errors-in-variables" problem in econometrics and its consequences for the least squares estimator.
3. Describe the properties of a good instrumental variable.
4. Discuss how the method of moments can be used to derive the least squares and instrumental variables estimators, paying particular attention to the assumptions upon which the derivations are based.
5. Explain why it is important for an instrumental variable to be highly correlated with the random explanatory variable for which it is an instrument.
6. Describe how instrumental variables estimation is carried out in the case of surplus instruments.
7. State the approximate large-sample distribution of the instrumental variables estimator for the simple linear regression model, and how it can be used for the construction of interval estimates and hypothesis tests.
8. Describe a test for the existence of contemporaneous correlation between the error term and the contemporaneous explanatory variables in a model, explaining the null and alternative hypotheses, and the consequences of rejecting the null hypothesis.

KEYWORDS

asymptotic properties
conditional expectation
endogenous variables
errors-in-variables
exogenous variables
first-stage regression
Hausman test
instrumental variable
instrumental variable estimator
just-identified
large sample properties
overidentified
population moments
random sampling
reduced-form
sample moments
sampling properties
simultaneous equations bias
surplus moment conditions
two-stage least squares estimation
weak instruments

In this chapter we reconsider the linear regression model. We will initially discuss the simple linear regression model, but our comments apply to the general model as well. The usual assumptions are SR1–SR6, given in Section 2.2.2. In Chapter 8, we relaxed the assumption $\text{var}(e_i|\mathbf{X}) = \sigma^2$ that the error variance is the same for all observations. In Chapter 9 we considered regressions with time-series data in which the assumption of serially uncorrelated errors, $\text{cov}(e_i, e_j|\mathbf{X}) = 0$, for $i \neq j$, cannot be maintained.

In this chapter, we relax the exogeneity assumption. When an explanatory variable is random, the properties of the least squares estimator depend on the characteristics of the independent variable x. The assumption of **strict exogeneity** is SR2 in the simple regression model, $E(e_i|\mathbf{x}) = 0$, and it is MR2 in the multiple regression model, $E(e_i|\mathbf{X}) = 0$. The mathematical form of this assumption is simple but the full meaning is complex. In Section 2.10.2, we gave common simple regression model examples when this assumption might fail. In these cases, with an explanatory variable that is **endogenous**, the usual least squares estimator does not have its desirable properties; it is not an unbiased estimator of the population parameters $\beta_1, \beta_2, \ldots$; it is not a consistent estimator of $\beta_1, \beta_2, \ldots$; tests and interval estimators do not have the anticipated properties, and even having large data samples will not cure the problems.

We review and discuss the properties of the least squares estimator with an endogenous explanatory variable in this chapter, and we suggest a new estimator, the **instrumental variables** estimator, that does have some desirable properties in large samples. The instrumental variables estimator is also called a **method of moments** estimator, and also the **two-stage least squares** estimator. We offer fair warning, however, that this area of econometrics is filled with practical and theoretical difficulties. Our search turns from finding an estimator that is "best" to one that is "adequate," and unfortunately producing convincing research applications requires knowledge, skill, and patience. In order for you to begin properly you should reread (right now!) Section 2.10 on the exogeneity concept and Section 5.7 on the large sample, or asymptotic, properties of the least squares estimator.

10.1 Least Squares Estimation with Endogenous Regressors

As our starting point, let us assume we are working with microeconomic, cross-sectional data obtained by **random sampling**. The standard assumptions for the simple regression model are RS1–RS6, which we repeat here for your convenience.

The Simple Linear Regression Model Under Random Sampling

RS1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

RS2: The data pairs (y_i, x_i) are statistically independent of all other data pairs and have the same joint distribution $f(y_i, x_i)$. They are independent and identically distributed (iid).

RS3: $E(e_i|x_i) = 0$ for $i = 1, \ldots, N$; x is contemporaneously, and strictly, exogenous.

RS4: The random error has constant conditional variance, $\text{var}(e_i|x_i) = \sigma^2$.

RS5: x_i takes at least two different values.

RS6: $e_i \sim N(0, \sigma^2)$

With random sampling, the ith and jth observations are statistically independent, so that the ith error e_i is statistically independent from the jth value of the explanatory variable, x_j. Thus, the

strict exogeneity assumption $E(e_i|x_1, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$ reduces to the simpler contemporaneous exogeneity assumption $E(e_i|x_i) = 0$.

Recall from Chapter 2 that the "gold standard" in research is a randomized controlled experiment. In an ideal (research) world, we would randomly assign x values (the treatment) and examine changes in outcomes y (the effect). If there is a systematic relationship between changes in x and changes in the outcome y, we can claim that changes in x **cause** changes in the outcome y. Any other random factors, "everything else" $= e$, that might affect the outcome are statistically independent of x. We can isolate, or identify, the effects of changes in x alone, and using regression analysis, we can estimate the causal effect $\Delta E(y_i|x_i)/\Delta x_i = \beta_2$.

The importance of the strict exogeneity assumption $E(e_i|x_i) = 0$ is that if it is true then "x is as good as randomly assigned." If $E(e_i|x_i) = 0$, then the best prediction of the random error e_i is simply zero. [See Appendix 4C for the details behind this statement.] There is no information contained in the values of x that helps us predict the random error. We can infer a causal relationship between y_i and x_i when there is covariation between them because variations in the random error e_i are uncorrelated with the variations in the explanatory variable x_i. It is just "as if" we had randomly assigned the treatments, x_i, to experimental subjects. Furthermore under RS1–RS6, the least squares estimators of β_1 and β_2 are the best linear unbiased estimators and the usual interval estimators and hypothesis tests work as they are expected to in samples of all sizes.

10.1.1 Large Sample Properties of the OLS Estimator

In Section 5.7, we introduced "large sample" or "asymptotic" analysis. With large samples of data, strict exogeneity is not required to identify and estimate a causal effect. All that we require is the simpler condition that the x values are uncorrelated with the random errors, e, and that the average of the random errors is zero. Econometricians, statisticians, and mathematicians aim to develop methods that work with as few strong assumptions as possible. We adopt that attitude and replace RS3, strict exogeneity, with

$$\text{RS3}^*: \; E(e_i) = 0 \text{ and } \text{cov}(x_i, e_i) = 0$$

Instead of contemporaneous exogeneity, we simply assume that the random error e_i and the explanatory variable value x_i are **contemporaneously uncorrelated**, which is a weaker condition than $E(e_i|x_i) = 0$. The term **contemporaneous** means "occurring at the same point in time" or, as in this case, occurring for the same cross-sectional observation subscript i. Explanatory variables like this, that are contemporaneously uncorrelated with the regression error, are simply said to be **exogenous**.

If we have obtained a random sample, then the selection of any person is statistically independent of the selection of any other person. Any randomly selected person's characteristics, such as education, income, ability, and race, are statistically independent of the characteristics of any other person selected. Because random sampling automatically implies zero correlation between the ith and jth observations, we only require that the ith value x_i be uncorrelated with e_i. The correlation between the ith error e_i and the jth value of the explanatory variable, x_j, is zero automatically because of random sampling.

Regression assumption RS3* says two things. First, in a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the population average of all unobservable characteristics, or variables omitted from the regression model, is zero, $E(e_i) = 0$. Second, in the population the correlation between the explanatory variable x_i and all the factors combined into the random error e_i is zero, or $\text{cov}(x_i, e_i) = 0$.

We can replace RS3 by RS3* because, if assumption RS3 is true, it follows that RS3* is true, that is, $E(e_i|x_i) = 0 \Rightarrow \text{cov}(x_i, e_i) = 0$ and $E(e_i|x_i) = 0 \Rightarrow E(e_i) = 0$. These relations are proven in Appendix 2G.1. Introducing assumption RS3* is convenient because it is a simpler notion of exogeneity, which is good. However, assumption RS3* is weaker than RS3 and under it we cannot show that the least squares estimator is unbiased, or that any of the other properties hold in small samples. What we can show is that the least squares estimators have

desirable **large sample properties**. Under assumptions RS1, RS2, RS3*, RS4, and RS5 the least squares estimators:

1. are consistent; that is, they converge in probability to the true parameter values as $N \to \infty$;
2. have approximate normal distributions in large samples, whether the random errors are normally distributed or not; and
3. provide interval estimators and test statistics that are valid if the sample is large.

In practice, this means that all the usual interpretations, intervals estimates, hypothesis tests, predictions, and prediction intervals are fine as long as our sample is large and RS1, RS2, RS3*, RS4, and RS5 hold. If samples are large, and if $\text{cov}(x_i, e_i) = 0$ and $E(e_i) = 0$, then it is "almost as good as" randomly assigning treatment values to x_i. We can estimate the population parameters $\beta_1, \beta_2, \ldots$ using the least squares estimator. If there is serial correlation or heteroskedasticity, then the robust standard error methods from Chapters 8 and 9 are fine as long as RS3* holds.

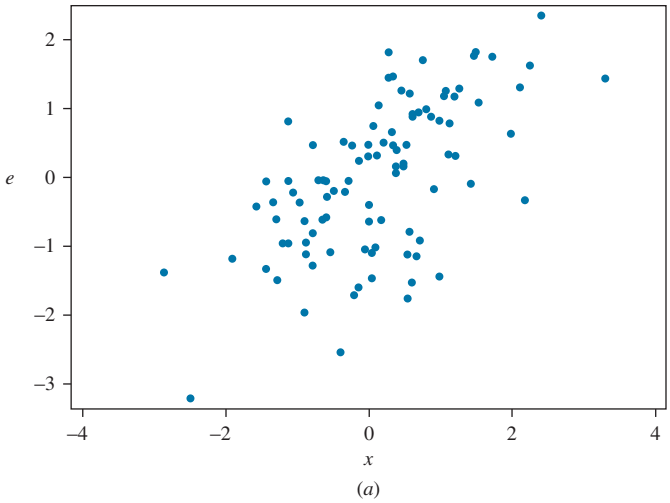
Remark

Do not fall into the trap of thinking "I'll just assume this, or that, if I want this or that result." It is true that access to large samples of data means not having to worry about the complexities of strict exogeneity. But what if you do not have access to large samples? Then statistical inference (estimation, hypothesis testing, and prediction) in small, or finite, samples is important. When the sample size N is not large, the **asymptotic properties** of estimators may be very misleading. Estimators that may be fine in large samples may suffer large biases in small samples. Estimates may appear statistically significant when they are not, and confidence intervals may be too narrow or too wide. If governments, or businesses, make decisions based on faulty inferences then we may suffer large economic or personal losses as a result. It is not just a game.

If assumption RS3* is *not* true, and in particular if $\text{cov}(x_i, e_i) \neq 0$ so that x_i and e_i are contemporaneously correlated, then the least squares estimators are inconsistent. They do not converge to the true parameter values even in very large samples. Furthermore, our usual hypothesis testing or interval estimation procedures are not valid. This means that estimating causal relationships using the least squares estimator when $\text{cov}(x_i, e_i) \neq 0$ may lead to incorrect inferences. When x_i is random, the relationship between x_i and e_i is a crucial factor when deciding whether least squares estimation, either OLS or GLS, is appropriate or not. If the error term e_i is correlated with x_i (or any x_{ik} in the multiple regression model) then the least squares estimator fails. In the next section we explain why correlation between x_i and e_i leads to the failure of the least squares estimator.

10.1.2 Why Least Squares Estimation Fails

In this section, we provide an intuitive explanation why the least squares estimator fails when $\text{cov}(x_i, e_i) \neq 0$. An algebraic proof is in the next section. The regression model **data generation process** adds a random error e_i to the systematic regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ to obtain the observed outcome y_i. In Figure 10.1(a), x_i and e_i values are positively correlated, violating the strict exogeneity assumption. In Figure 10.1(b), the positively sloped regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, which is the object of our analysis, is the solid line. For each value of x_i, the y_i data values, $y_i = \beta_1 + \beta_2 x_i + e_i$, are the sum of the systematic portion $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ and a random error e_i. The data pairs (y_i, x_i) are the dots in Figure 10.1(b). As you see, the true regression function does not pass through the middle of the data in this case and that is because of the correlation between x_i and e_i. The y_i values for larger x_i values tend to have positive errors, $e_i > 0$. The y_i values for smaller x_i values have negative errors, $e_i < 0$. In this case, we can use



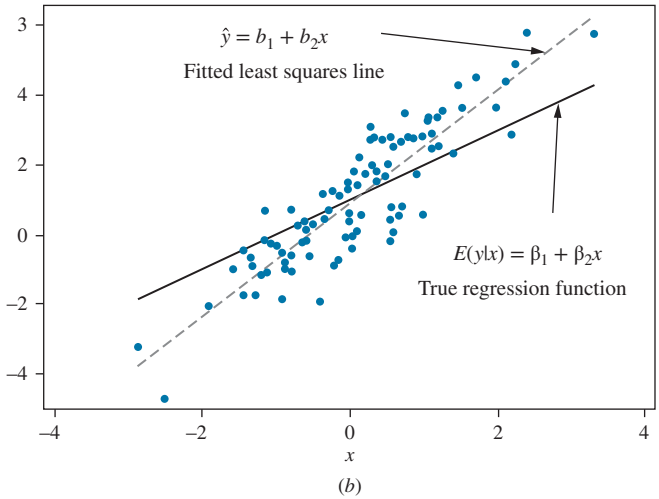


FIGURE 10.1 **(a) Correlated x and e. (b) Plot of data, true and fitted regression functions.**

information provided by the x_i values to provide a better prediction of the random error e_i than simply zero.

Least squares estimation leads to a fitted line passing through the middle of the data, shown as a dashed line in Figure 10.1(b). The slope of the fitted line $\left(\text{the estimate } b_2\right)$ overestimates the true slope of the regression function, $\beta_2 > 0$. The least squares estimator attributes all variation in y_i to variation in x_i. When x_i and e_i are correlated, the variation in y_i comes from two sources: changes in x_i and changes in e_i, and in our example these changes have a positive correlation. If we think about the effect of changes in x_i and e_i on y_i we have

$$\begin{array}{ccccc} \Delta y_i & = & \beta_2 \Delta x_i & + & \Delta e_i \\ (+) & & (+) & & (+) \end{array}$$

If x_i and e_i are positively correlated and $\beta_2 > 0$, increases in x_i and e_i combine to increase y_i. In the least squares estimation process, all the change (increase) in y_i is attributed to the effect of the change (increase) in x_i, and thus the least squares estimator will overestimate β_2.

Throughout this Chapter, we use the relation between wages and years of education as an example. In this case, the omitted variable "intelligence," or ability, is in the regression error, and it is likely to be positively correlated with the years of education a person receives, with more

intelligent individuals usually choosing to obtain more years of education. When regressing wage on years of education, the least squares estimator attributes increases in wages to increases in education. The effect of education is overstated because some of the increase in wages is also due to higher intelligence.

The statistical consequence of a contemporaneous correlation between x_i and e_i is that the least squares estimator is biased, and this bias will not disappear no matter how large the sample is. Consequently, the least squares estimator is **inconsistent** when there is contemporaneous correlation between x_i and e_i.

Remark

If x_i is endogenous the least squares estimator still is a useful **predictive** tool. In Figure 10.1(b) the least squares fitted line fits the data well. Given a value x_0 we can predict y_0 using the fitted line. What we cannot do is interpret the slope of the line as a causal effect.

10.1.3 Proving the Inconsistency of OLS

Let us prove that the least squares estimator is not consistent when $\text{cov}(x_i, e_i) \neq 0$. Our regression model is $y_i = \beta_1 + \beta_2 x_i + e_i$. Continue to assume that $E(e_i) = 0$, so that $E(y_i) = \beta_1 + \beta_2 E(x_i)$. Then,

- Subtract this expectation from the original equation,

$$y_i - E(y_i) = \beta_2\left[x_i - E(x_i)\right] + e_i$$

- Multiply both sides by $x_i - E(x_i)$

$$\left[x_i - E(x_i)\right]\left[y_i - E(y_i)\right] = \beta_2\left[x_i - E(x_i)\right]^2 + \left[x_i - E(x_i)\right]e_i$$

- Take expected values of both sides

$$E\left[x_i - E(x_i)\right]\left[y_i - E(y_i)\right] = \beta_2 E\left[x_i - E(x_i)\right]^2 + E\left\{\left[x_i - E(x_i)\right]e_i\right\},$$

or

$$\text{cov}(x_i, y_i) = \beta_2 \text{var}(x_i) + \text{cov}(x_i, e_i)$$

- Solve for β_2

$$\beta_2 = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)} - \frac{\text{cov}(x_i, e_i)}{\text{var}(x_i)}$$

This equation is the basis for showing when the least squares estimator is consistent, and when it is not.

If we can assume that $\text{cov}(x_i, e_i) = 0$, then

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

We drop the "i" subscript because we are randomly sampling from a population, and the data pairs are not only independently distributed but identically distributed, with the same joint *pdf* $f(x_i, y_i)$, and thus $\text{cov}(x_i, y_i) = \text{cov}(x, y)$ and $\text{var}(x_i) = \text{var}(x)$. The least squares estimator is

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/(N-1)}{\sum(x_i - \bar{x})^2/(N-1)} = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)}$$

This shows that the least squares estimator b_2 is the sample analog of the population relationship, $\beta_2 = \text{cov}(x, y)/\text{var}(x)$. The sample variance and covariance converge to the true variance and covariance as the sample size N increases, using the Law of Large Numbers introduced in Section 10.3.1, so that the least squares estimator converges to β_2. That is, if $\text{cov}(x_i, e_i) = 0$, then

$$b_2 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} \rightarrow \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2$$

showing that the least squares estimator is consistent.

On the other hand, if x_i and e_i are correlated, then

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)} - \frac{\text{cov}(x, e)}{\text{var}(x)}$$

The least squares estimator now converges to

$$b_2 \rightarrow \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2 + \frac{\text{cov}(x, e)}{\text{var}(x)} \neq \beta_2$$

In this case, b_2 is an inconsistent estimator of β_2 and the amount of bias that exists even asymptotically, when samples can be assumed to be large, is $\text{cov}(x, e)/\text{var}(x)$. The direction of the bias depends on the sign of the covariance between x_i and e_i. If factors in the error are positively correlated with the explanatory variable x, then the least squares estimator will overestimate the true parameter. If factors in the error are negatively correlated with the explanatory variable x, then the least squares estimator will underestimate the true parameter.

In the following section, we describe some common situations in which there is correlation between x_i and e_i causing the least squares estimator to fail.

10.2 Cases in Which x and e are Contemporaneously Correlated

There are several common situations in which the least squares estimator fails due to the presence of a contemporaneous correlation between an explanatory variable and the error term. When an explanatory variable and an error term are contemporaneously correlated, the explanatory variable is said to be **endogenous**. This term comes from simultaneous equations models, which we will consider in Chapter 11, and means "determined within the system." When an explanatory variable is contemporaneously correlated with the regression error one is said to have an "endogeneity problem."

10.2.1 Measurement Error

The **errors-in-variables** problem occurs when an explanatory variable is measured with error. If we measure an explanatory variable with error, then it is correlated with the error term, and the least squares estimator is inconsistent. As an illustration, consider the following important example. Let us assume that an individual's personal saving is based on their "permanent" or long-run income. Let y_i = annual savings of the ith person and let x_i^* = the permanent annual income of the ith person. A simple regression model representing this relationship is

$$y_i = \beta_1 + \beta_2 x_i^* + v_i \tag{10.1}$$

We have asterisked (*) the permanent income variable because it is difficult, if not impossible, to observe. For the purposes of a regression, suppose that we attempt to measure permanent income

using x_i = current income. Current income is a measure of permanent income, but it does not measure permanent income exactly. To capture this measurement error specify that

$$x_i = x_i^* + u_i \tag{10.2}$$

where u_i is a random disturbance, with mean 0 and variance σ_u^2. With this statement, we are admitting that observed current income only approximates permanent income, and consequently that we have measured permanent income with error. Furthermore, assume that the measurement error u_i is independent of the regression error v_i. When we use x_i in the regression in place of x_i^*, we do so by replacement, that is, substitute $x_i^* = x_i - u_i$ into (10.1) to obtain

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i^* + v_i = \beta_1 + \beta_2(x_i - u_i) + v_i = \beta_1 + \beta_2 x_i + (v_i - \beta_2 u_i) \\ &= \beta_1 + \beta_2 x_i + e_i \end{aligned} \tag{10.3}$$

In order to estimate (10.3) by OLS, we must determine whether or not x_i is contemporaneously uncorrelated with the random error e_i. The covariance between these two random variables, using the fact that $E(e_i) = 0$ and assuming that x_i^* is exogenous in (10.1), so that $E(x_i^* v_i) = 0$, is

$$\begin{aligned} \operatorname{cov}(x_i, e_i) &= E(x_i e_i) = E\left[(x_i^* + u_i)(v_i - \beta_2 u_i)\right] \\ &= E(-\beta_2 u_i^2) = -\beta_2 \sigma_u^2 \neq 0 \end{aligned} \tag{10.4}$$

The least squares estimator b_2 is an *inconsistent* estimator of β_2 in (10.3) because of the correlation between the explanatory variable x_i and the error term e_i. Consequently, b_2 does not converge to β_2 in large samples. Furthermore, in large or small samples, b_2 is *not* approximately normal with mean β_2 and variance $\operatorname{var}(b_2) = \sigma^2 / \sum(x_i - \bar{x})^2$. When ordinary least squares fails in this way, is there another estimation approach that works? The answer is yes, as we will see in Section 10.3.

Note that in equation (10.4), if $\beta_2 > 0$, there is a negative correlation between x_i and the random error e_i. The least squares estimator will underestimate β_2 and in the literature devoted to measurement error this is called **attenuation bias**. This is a logical result of using $x_i = x_i^* + u_i$. Imagine that the measurement error u_i is very large relative to x_i^*. Then x_i becomes more like a completely random number and there will be little association between y_i and x_i in the data, so that b_2 will be near zero.

10.2.2 Simultaneous Equations Bias

Another situation in which an explanatory variable is correlated with the regression error term arises in simultaneous equations models. While this terminology may not sound familiar, students of economics deal with such models from their earliest introduction to supply and demand. Recall that in a competitive market the prices and quantities of goods are determined jointly by the forces of supply and demand. Thus, if P_i = equilibrium price and Q_i = equilibrium quantity, we can say that P_i and Q_i are endogenous, because they are jointly determined within a simultaneous system of two equations, one equation for the supply curve and one equation for the demand curve. Suppose that we write down the relation

$$Q_i = \beta_1 + \beta_2 P_i + e_i \tag{10.5}$$

We know that changes in price affect the quantities supplied and demanded. But it is also true that changes in quantities supplied and demanded lead to changes in prices. There is a feedback relationship between P_i and Q_i. Because of this feedback, which results because price and quantity are jointly, or simultaneously, determined, we can show that $\operatorname{cov}(P_i, e_i) \neq 0$. The least squares estimation procedure will fail if applied to (10.5) because of the endogeneity problem, and the resulting bias (and inconsistency) is called **simultaneous equations bias**. Supply and demand models permeate economic analysis, and we will treat simultaneous equations models fully in Chapter 11.

10.2.3 Lagged-Dependent Variable Models with Serial Correlation

In Chapter 9, we introduced dynamic models with stationary variables. One way to make models dynamic is to introduce a lagged dependent variable into the right-hand side of an equation, That is, $y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$. The lagged variable y_{t-1} is a random regressor, but as long as it is uncorrelated with the error term e_t then the least squares estimator is consistent. However, it is possible when specifying a dynamic model that the errors will be serially correlated. If the errors e_t follow the AR(1) process $e_t = \rho e_{t-1} + v_t$, then we can see that the lagged dependent variable y_{t-1} must be correlated with the error term e_t, because y_{t-1} depends directly on e_{t-1}, and e_{t-1} directly affects the value of e_t. If $\rho \neq 0$, there will be a correlation between y_{t-1} and e_t. In this case, the OLS estimator applied to the lagged dependent variable model will be biased and inconsistent. Thus, it is very important to test for the presence of serial correlation in models with lagged dependent variables on the right-hand side (see Sections 9.4 and 9.5).

10.2.4 Omitted Variables

When an omitted variable is correlated with an included explanatory variable, then the regression error will be correlated with the explanatory variable. We introduced this idea in Section 6.3.1. A classic example is from labor economics. A person's wage is determined in part by their level of education. Let us specify a log-linear regression model explaining observed hourly wage as

$$\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + e_i \tag{10.6}$$

with $EDUC_i$ = years of education and $EXPER_i$ = years of work experience. What else affects wages? What is omitted from the model? This thought experiment should be carried out each time a regression model is formulated. There are several factors we might think of, such as labor market conditions, region of the country, and union membership. However, labor economists are most concerned about the omission of a variable measuring ability. It is logical that a person's ability, intelligence and industriousness may affect the quality of their work and their wage. These variables are components of the random error e_i, since we usually have no measure for them. The problem is not only that ability might affect wages but more able individuals may also spend more years in school, causing a positive correlation between the error terms e_i and $EDUC_i$, so that $\text{cov}(EDUC_i, e_i) > 0$. If this is true, then we can expect that the least squares estimator of the returns to another year of education will be positively biased, $E(b_2) > \beta_2$, and inconsistent, meaning that the bias will not disappear even in very large samples.

EXAMPLE 10.1 | Least Squares Estimation of a Wage Equation

We will use the data on married women in the data file *mroz* to estimate the wage model in (10.6). Using the $N = 428$ women in the sample who are in the labor force, the least squares estimates and their standard errors are

$$\begin{aligned} \ln(WAGE) &= -0.5220 + 0.1075 \times EDUC \\ \text{(se)} \qquad & \quad (0.1986) \quad (0.0141) \\ &+ 0.0416 \times EXPER - 0.0008 \times EXPER^2 \\ & \quad (0.0132) \qquad\quad (0.0004) \end{aligned}$$

We estimate that an additional year of education increases wages by approximately 10.75%, holding everything else constant. If ability has a positive effect on wages, then this estimate is overstated, as the contribution of ability is attributed to the education variable.

The social and policy importance of the estimate 0.1075 can hardly be exaggerated. Countries invest a large portion of tax revenue to improve education. Why? Spending on education is an investment, and like any other investment investors (taxpaying citizens) expect a rate of return that is competitive with rates of returns for alternative projects. Based on the estimated equation above, additional years of schooling are estimated to increase wages by 10.75%, holding other factors fixed, meaning that individuals are

more likely to be self-sufficient, enjoy a good quality of life, not requiring welfare or public health assistance, and less likely to engage in crime. Suppose, however, that 10.75% overestimates the returns to education for wage income. We might re-evaluate the investment in education and perhaps decide to spend tax dollars on bridges or parks instead of schools. Evaluating the social rate of return to education is a social policy problem. Regression estimates such as those above play heavily into the calculation. Consequently we must do all that we can, as econometricians, to obtain estimates using the best methods. In the next section we begin our examination of alternative estimation methods for models in which regression errors are correlated with regression variables.

10.3 Estimators Based on the Method of Moments

In the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, when x_i is random and $\text{cov}(x_i, e_i) \neq 0$, the least squares estimators are biased and inconsistent, with none of their usual nice properties holding. When faced with such a situation we must consider alternative estimation procedures. In this section we discuss the "method of moments" principle of estimation, which is an alternative to the least squares estimation principle. When all the usual assumptions of the linear model hold, the method of moments leads us to the least squares estimator. If x_i is random and correlated with the error term, the method of moments leads us to an alternative, called instrumental variables estimation or two-stage least squares estimation, that will work in large samples.

10.3.1 Method of Moments Estimation of a Population Mean and Variance

Let us begin with a simple case. The kth moment of a random variable Y is the expected value of the random variable raised to the kth power. That is,

$$E(Y^k) = \mu_k = k\text{th moment of } Y \tag{10.7}$$

The **Law of Large Numbers (LLN)** is a famous theorem. One version says: if $X_1, X_2, \ldots, X_N$ is a random sample from a population, and if $E(X_i) = \mu < \infty$ and $\text{var}(X_i) = \sigma^2 < \infty$, then the sample mean $\overline{X} = \sum X_i/N$ converges (in probability) to the expected value (population mean) μ as the sample size N increases. In this case, $\overline{X}$ is said to be a consistent estimator of μ. It is useful to remember that in most situations **sample moments** are consistent estimators of **population moments**.

We can apply the law of large numbers to obtain a consistent estimator of $E(Y^k) = \mu_k$ by letting $X_i = Y_i^k$ and $E(X_i) = \mu = E(Y_i^k) = \mu_k$. Then, assuming that $\text{var}(Y_i^k) = \sigma_k^2 < \infty$, a consistent estimator of the population moment $E(Y^k) = \mu_k$ is the corresponding sample moment

$$\widehat{E(Y^k)} = \hat{\mu}_k = k\text{th sample moment of } Y = \sum Y_i^k/N \tag{10.8}$$

The **method of moments** estimation procedure equates m population moments to m sample moments to estimate m unknown parameters. As an example, let Y be a random variable with mean $E(Y) = \mu$ and variance, given in the Probability Primer, equation (P.13):

$$\text{var}(Y) = \sigma^2 = E(Y - \mu)^2 = E(Y^2) - \mu^2 \tag{10.9}$$

In order to estimate the two population parameters μ and σ^2, we must equate two population moments to two sample moments. Let $Y_1, Y_2, \ldots, Y_N$ be a random sample from the population.

The first two population and sample moments of Y are

$$\begin{array}{ll} \text{Population moments} & \text{sample moments} \\ E(Y) = \mu_1 = \mu & \hat{\mu} = \sum Y_i / N \\ E(Y^2) = \mu_2 & \hat{\mu}_2 = \sum Y_i^2 / N \end{array} \tag{10.10}$$

Note that for the first population moment μ_1, it is customary to drop the subscript and use μ to denote the population mean of Y. With these two moments, we can solve for the unknown mean and variance parameters. Equate the first sample moment in (10.10) to the first population moment to obtain an estimate of the population mean,

$$\hat{\mu} = \sum Y_i / N = \overline{Y} \tag{10.11}$$

Then use (10.9), replacing the second population moment in (10.10) by its sample value and replacing first moment μ by (10.11)

$$\tilde{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{\sum Y_i^2}{N} - \overline{Y}^2 = \frac{\sum Y_i^2 - N\overline{Y}^2}{N} = \frac{\sum \left(Y_i - \overline{Y}\right)^2}{N} \tag{10.12}$$

The method of moments leads us to the sample mean as an estimator of the population mean. The method of moments estimator of the variance has N in its denominator, rather than the usual $N-1$, so it is not exactly the sample variance we are used to. But in large samples this will not make much difference. In general, method of moments estimators are consistent, and converge to the true parameter values in large samples, but there is no guarantee that they are "best" in any sense.

10.3.2 Method of Moments Estimation in the Simple Regression Model

The definition of a "moment" can be extended to more general situations. Assumption RS3* states that $E(e_i) = 0$ and $\text{cov}(x_i, e_i) = E(x_i e_i) = 0$. Using these two equations, we can derive the OLS estimator by using the method of moments approach. In the linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the two moment conditions $E(e_i) = 0$ and $E(x_i e_i) = 0$ imply

$$E(e_i) = 0 \Rightarrow E(y_i - \beta_1 - \beta_2 x_i) = 0 \tag{10.13}$$

and

$$E(x_i e_i) = 0 \Rightarrow E\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0 \tag{10.14}$$

Equations (10.13) and (10.14) are population moment conditions. The Law of Large Numbers says that under random sampling, sample moments converge to population moments, so

$$\frac{1}{N}\sum(y_i - \beta_1 - \beta_2 x_i) \xrightarrow{p} E(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] \xrightarrow{p} E\left[x_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0$$

Setting the two sample moment conditions to zero and replacing the unknown parameters β_1 and β_2 by their estimators b_1 and b_2, we have two equations and two unknowns

$$\frac{1}{N}\sum(y_i - b_1 - b_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[x_i(y_i - b_1 - b_2 x_i)\right] = 0$$

Multiplying these two equations by N we have the two **normal equations** (2A.3) and (2A.4) given in Appendix 2A, and solving them yields the least squares estimators,

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

What we have shown is that under the weaker assumptions, $E(e_i) = 0$ and zero contemporaneous covariance between x_i and e_i, $\text{cov}(x_i, e_i) = E(x_i e_i) = 0$, we can derive the OLS estimators for the simple linear regression model using the method of moments approach. Further, as we have discussed in Section 5.7, the OLS estimators are **consistent estimators** in this case, and have their usual properties in large samples.

10.3.3 Instrumental Variables Estimation in the Simple Regression Model

Problems for least squares estimation arise when x_i is random and contemporaneously correlated with the random error e_i, so that $\text{cov}(x_i, e_i) = E(x_i e_i) \neq 0$. In this case x_i is **endogenous**. As we have discussed in Sections 5.7 and 6.3, and Appendix 6B, the OLS estimator is biased and **inconsistent** when an explanatory variable is endogenous. Also, in the method of moments context, endogeneity makes the moment condition in equation (10.14) invalid.

What are we to do? The method of moments approach gives us an insight into an alternative. Suppose that there is another variable, z_i, with the following properties:

Characteristics of a Good Instrumental Variable

IV1: z_i does not have a direct effect on y_i, and thus it does not belong on the right-hand side of the model $y_i = \beta_1 + \beta_2 x_i + e_i$ as an explanatory variable.

IV2: z_i is not contemporaneously correlated with the regression error term e_i, so that $\text{cov}(z_i, e_i) = E(z_i e_i) = 0$. Variables with the property $\text{cov}(z_i, e_i) = E(z_i e_i) = 0$ are said to be **exogenous**.

IV3: z_i is strongly (or at least not weakly) correlated with x_i, the endogenous explanatory variable.

A variable z_i with these properties is called an **instrumental variable**. This terminology arises because while z does not have a direct effect on y, having it will allow us to estimate the relationship between x and y. It is a *tool*, or **instrument**, that we are using to achieve our objective.

If such a variable z exists, then we can use it to form a moment condition to replace (10.14), that is,

$$E(z_i e_i) = 0 \Rightarrow E\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0 \tag{10.15}$$

Then we can use the two moment equations (10.13) and (10.15) to obtain estimates of β_1 and β_2. Again appealing to the Law of Large numbers, we can assert that sample moments converge to population moments. Therefore,

$$\frac{1}{N}\sum(y_i - \beta_1 - \beta_2 x_i) \xrightarrow{p} E(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] \xrightarrow{p} E\left[z_i(y_i - \beta_1 - \beta_2 x_i)\right] = 0$$

Assuming we have a sufficiently large sample, we set the sample moments to zero, yielding the two sample moment conditions

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$
$$\frac{1}{N}\sum z_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0 \tag{10.16}$$

Solving these equations leads us to method of moments estimators, which in economics are usually called the **instrumental variable (IV) estimators**,

$$\hat{\beta}_2 = \frac{N\sum z_i y_i - \sum z_i \sum y_i}{N\sum z_i x_i - \sum z_i \sum x_i} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \tag{10.17}$$

We introduce the notation $\hat{\beta}_1$ and $\hat{\beta}_2$ for the instrumental variables estimators to differentiate them from the OLS estimators b_1 and b_2. If properties IV1, IV2, and IV3 hold, then these new estimators are **consistent**, they converge to the true parameter values as the sample size $N \to \infty$. Also, they have approximate normal distributions in large samples, which we denote by "$\overset{a}{\sim}$". For the simple regression model

$$\hat{\beta}_2 \overset{a}{\sim} N\left[\beta_2, \widehat{\text{var}}\left(\hat{\beta}_2\right)\right]$$

where the estimated variance is

$$\widehat{\text{var}}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2_{IV}\sum(z_i - \bar{z})^2}{\left[\sum(z_i - \bar{z})(x_i - \bar{x})\right]^2} \tag{10.18a}$$

The IV estimator of the error variance σ^2 is

$$\hat{\sigma}^2_{IV} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N - 2} \tag{10.18b}$$

10.3.4 The Importance of Using Strong Instruments

When working with instrumental variables, a constantly repeated question is "How strong are the instruments?" What is a strong instrument? We will develop a full answer to that question in this chapter, but initially, we define a strong instrument z as one that is highly correlated with the endogenous variable x. To show why this definition is useful, apply a bit of algebra to the expression for the variance $\text{var}\left(\hat{\beta}_2\right)$ in equation (10.18a) to obtain an informative equivalent expression.

$$\widehat{\text{var}}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2_{IV}\sum(z_i - \bar{z})^2}{\left[\sum(z_i - \bar{z})(x_i - \bar{x})\right]^2}$$
$$= \frac{\hat{\sigma}^2_{IV}}{\left\{\dfrac{\left[\sum(z_i - \bar{z})(x_i - \bar{x})\right]^2 \Big/ (N-1)}{\sum(z_i - \bar{z})^2 \sum(x_i - \bar{x})^2 \Big/ (N-1)}\right\}\sum(x_i - \bar{x})^2}$$
$$= \frac{\hat{\sigma}^2_{IV}}{r^2_{zx}\sum(x_i - \bar{x})^2}$$

We simply multiplied and divided by $\sum(x_i - \bar{x})^2$ and by $(N-1)$ in the middle equation and did some rearranging. The final expression tells us about the precision of estimation of the coefficient of the endogenous variable. As was the case with the OLS estimator, the variance of $\hat{\beta}_2$ depends on the variation in the explanatory variable about its mean, $\sum(x_i - \bar{x})^2$, and the estimated variance of the error term $\hat{\sigma}^2_{IV}$. Those components are familiar to you. What is new is that the denominator also includes the squared sample correlation r_{zx} between the instrumental variable z and the endogenous variable x. The larger the magnitude of the sample correlation $|r_{zx}|$ the smaller the estimated variance of the IV estimator, and vice versa. When $|r_{zx}|$ is large, the instrumental variable is strong. Stronger instrumental variables lead to smaller estimated variances, smaller standard errors, narrower interval estimates, and generally more precise statistical inference. It is important to choose strong instrumental variables.

To illustrate and make the point about instrument strength dramatic, suppose $\text{cov}(x_i, e_i) = 0$, so that both the OLS and IV estimators are consistent. Comparing the estimated variance of the two estimators, the ratio of the estimated variance of the IV estimator to that of the OLS estimator is

$$\frac{\widehat{\text{var}}(\hat{\beta}_2)}{\widehat{\text{var}}(b_2)} = \frac{\hat{\sigma}^2_{IV}}{\dfrac{r^2_{zx}\sum(x_i - \bar{x})^2}{\hat{\sigma}^2}} = \frac{\hat{\sigma}^2_{IV}/\hat{\sigma}^2}{r^2_{zx}} \simeq \frac{1}{r^2_{zx}}$$

The final approximation uses the fact that if $\text{cov}(x_i, e_i) = 0$, then in large samples the two estimators of σ^2 will converge to the same value so that $\hat{\sigma}^2_{IV}/\hat{\sigma}^2 \simeq 1$. The squared correlation $r^2_{zx} < 1$ and thus we anticipate that the variance estimate for the IV estimator will be larger than the variance estimate for the OLS estimator. The IV estimator is less *efficient* than the OLS estimator, meaning that it makes less efficient use of sample data to estimate the unknown parameters.

We prefer the more efficient consistent estimator because it has a smaller standard error, leading to narrower interval estimates, making statistical inferences more precise. The ratio of standard errors is $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 1/|r_{zx}|$. If the correlation $r_{zx} = 0.5$, then $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 2$, the estimated standard error of the IV estimator is two times as large as the standard error of the OLS estimator. If $r_{zx} = 0.1$, then $\text{se}(\hat{\beta}_2)/\text{se}(b_2) \simeq 10$, the estimated standard error of the IV estimator is *10* times as large as the standard error of the OLS estimator.

To put some meat on these bones, recall that in large samples a 95% interval estimate is approximately "estimate $\pm$ 2(standard error)." For the sake of illustration, suppose $b_2 \simeq \hat{\beta}_2 = 5$ and $\text{se}(b_2) = 1$, then the 95% interval estimate using the OLS estimator is $5 \pm 2(1)$ or $[3, 7]$. If $r_{zx} = 0.5$, then the interval estimate based on the IV estimator is $5 \pm 2(2)$ or $[1, 9]$. If $r_{zx} = 0.1$, then the interval estimate based on the IV estimator is $5 \pm 2(10)$ or $[-15, 25]$. This shocking difference will remind you not to use the IV estimator unless you have to. If you do have to use IV estimation, then you must search for a strong instrumental variable, one that is highly correlated with the endogenous x.

10.3.5 Proving the Consistency of the IV Estimator

The demonstration that the instrumental variables estimator is consistent follows the logic used in Section 10.1.3. The IV estimator of β_2 in (10.17) is

$$\hat{\beta}_2 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})/(N-1)}{\sum(z_i - \bar{z})(x_i - \bar{x})/(N-1)} = \frac{\widehat{\text{cov}}(z, y)}{\widehat{\text{cov}}(z, x)}$$

The sample covariance converges to the true covariance in large samples, so we can say

$$\hat{\beta}_2 \to \frac{\text{cov}(z, y)}{\text{cov}(z, x)}$$

If the instrumental variable z is not correlated with x in either the sample data or in the population, then the **instrumental variable estimator** fails. Having z and x uncorrelated in the sample data would mean a zero in the denominator of $\hat{\beta}_2$. Having z and x uncorrelated in the population means $\hat{\beta}_2$ would not converge in large samples. Thus for an instrumental variable to be valid, it must be uncorrelated with the error term e but correlated with the explanatory variable x.

Now, following the same steps as in Section 10.1.3, we obtain

$$\beta_2 = \frac{\text{cov}(z, y)}{\text{cov}(z, x)} - \frac{\text{cov}(z, e)}{\text{cov}(z, x)}$$

If we can assume that that $\text{cov}(z_i, e_i) = 0$, a condition we imposed on the choice of the instrumental variable z_i, then the instrumental variables estimator $\hat{\beta}_2$ converges in large samples to β_2,

$$\hat{\beta}_2 \to \frac{\text{cov}(z, y)}{\text{cov}(z, x)} = \beta_2$$

Thus, if $\text{cov}(z_i, e_i) = 0$ and $\text{cov}(z_i, x_i) \neq 0$, then the instrumental variable estimator of β_2 is consistent, in a situation in which the OLS estimator is not consistent due to correlation between x_i and e_i.

EXAMPLE 10.2 | IV Estimation of a Simple Wage Equation

To illustrate the instrumental variables estimation method in a simple regression consider a simplified version of the model used in Example 10.1, $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$. Using the data file *mroz* on $N = 428$ married women, the OLS estimates are

$$\widehat{\ln(WAGE)} = -0.1852 + 0.1086 EDUC$$
$$\text{(se)} \qquad (0.1852) \quad (0.0144)$$

The estimated rate of return to education is approximately 10.86%, and $t = 7.55$ indicates that the estimated coefficient is significantly different from zero at even the 1% level of significance. If *EDUC* is endogenous, and correlated with the random error e, then OLS estimation may lead to incorrect inferences. We anticipate that *EDUC* is positively correlated with the omitted variable "ability," meaning that the estimated rate of return 10.86% may overstate the true value.

What might we use as an instrumental variable? One proposal is to use mother's years of education, *MOTHEREDUC*, as an instrument. Does this qualify? In Section 10.3.3, we listed three criteria for an instrumental variable. First, does this variable have a direct effect on the dependent variable? Does it belong in the equation? Mother's education should not play any direct role in the determination of a daughter's wage, so this seems fine. Second, the instrument should not be contemporaneously correlated with the random error, e. Is a mother's education correlated with the omitted variable, her daughter's ability? This is more difficult. Ability includes many attributes, including intelligence, creativity, perseverance, and industriousness to name a few. Some portion of these character traits may be passed into our genetic makeup from our parents. We dodge the scientific debate on this issue and assume that a mother's years of education are uncorrelated with her daughter's ability. Third, the instrument should be highly correlated with the endogenous variable. This we can check! For the 428 women in the sample the correlation between mother's education and daughter's education is 0.3870. This is not very large, but it is not very small either.

The instrumental variables estimates are

$$\widehat{\ln(WAGE)} = 0.7022 + 0.0385 EDUC$$
$$\text{(se)} \qquad (0.4851) \quad (0.0382)$$

The IV estimate of the rate of return to education is 3.85%, one-third of the OLS estimate. The standard error is about 2.65 times larger than the OLS standard error, which is very close to what we reasoned that the ratio might be when both estimators are consistent,

$$\text{se}\left(\hat{\beta}_2\right) \Big/ \text{se}\left(b_2\right) = 0.0382/0.0144 = 2.65 \simeq 1/r_{zx}$$
$$= 1/0.3807 = 2.58$$

10.3.6 IV Estimation Using Two-Stage Least Squares (2SLS)

We can obtain the instrumental variables estimates by another type of calculation, one that will help us extend the IV estimation idea to more general situations. The method called **two-stage**

least squares uses two least squares regressions to calculate the IV estimates. The **first-stage equation** has a dependent variable that is the endogenous regressor x, and the independent variable z, the instrumental variable. That is, the first-stage equation is

$$x = \gamma_1 + \theta_1 z + v$$

where γ_1 is an intercept parameter, θ_1 is a slope parameter, and v is an error term. The steps in 2SLS are as follows:

1. Estimate the first-stage equation by OLS and obtain the fitted value, $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$.
2. In the **second stage**, replace the endogenous variable x in the simple regression $y = \beta_1 + \beta_2 x + e$ with $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$ and then apply OLS estimation to $y = \beta_1 + \beta_2 \hat{x} + e^*$.

The OLS estimates of β_1 and β_2 from the second-stage regression are identically equal to the IV estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Furthermore, the estimated variances and covariances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS formulas with $\hat{\sigma}^2_{IV} = \sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2 \Big/ (N-2)$ replacing the usual estimate of σ^2 and using the fact that $\bar{\hat{x}} = \bar{x}$,

$$\widehat{\text{var}}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2_{IV}}{\sum\left(\hat{x}_i - \bar{x}\right)^2} \tag{10.19}$$

This variance estimate is numerically identical to the previous expression in equation (10.18a). If (10.19) is not used, the second-stage OLS regression computes the variance incorrectly, because OLS software will use

$$\hat{\sigma}^2_{WRONG} = \sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i\right)^2 \Big/ (N-2)$$

putting $\hat{x}_i$ in place of x_i. Always use software designed for IV/2SLS as it will carry out the correct calculation.

EXAMPLE 10.3 | 2SLS Estimation of a Simple Wage Equation

To illustrate the two-stage least squares equivalent of instrumental variables estimation, we estimate the first-stage equation, a regression of the endogenous variable *EDUC* on the instrumental variable *MOTHEREDUC*

$$\widehat{EDUC} = 10.1145 + 0.2674 MOTHEREDUC$$
$$\text{(se)} \qquad (0.3109)\ \ (0.0309)$$

In order for *MOTHEREDUC* to be a strong instrumental variable it must be strongly correlated with *EDUC*. Another way to say this is that *MOTHEREDUC* should be strongly significant in this first-stage equation, and it is. The t-value is 8.66, so the coefficient is significantly different from zero at the 1% level. We will say much more about this approach in Section 10.3.9.

In the second-stage equation, we regress ln(*WAGE*) on the fitted value from the first-stage equation,

$$\widehat{\ln(WAGE)} = 0.7021 + 0.0385 \widehat{EDUC}$$
$$\text{(incorrect se)}\ (0.5021)\ \ (0.0396)$$

The coefficient estimates are the same as in Example 10.2, but note that the standard errors produced by this second OLS estimation are not the same as in Example 10.2. They are **incorrect** because they use $\hat{\sigma}^2_{WRONG}$.

10.3.7 Using Surplus Moment Conditions

The reason for introducing two-stage least squares is that it is an easy way to use extra, additional, instrumental variables. In a simple regression, we need only one instrumental variable, yielding two moment conditions like (10.16), which we solve for the two unknown model parameters.

Sometimes, however, we have more instrumental variables than are necessary. Suppose we have two good instruments, z_1 and z_2 that satisfy conditions IV1–IV3. Compared to (10.16) we have the additional moment condition

$$E(z_2 e) = E\left[z_2(y - \beta_1 - \beta_2 x)\right] = 0$$

There are now three sample moment conditions:

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum z_{i1}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum z_{i2}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

We have three equations with only two unknowns. There are no solutions satisfying all three equations. We could simply throw away one of the conditions (instruments) and use the remaining two to solve for the unknowns. A better solution is to use all the available instruments by combining them. It can be proved that the best way of combining instruments is using the two-stage least squares idea. In the simple regression $y = \beta_1 + \beta_2 x + e$, if x is endogenous, and we have two instruments, z_1 and z_2, the first-stage equation becomes

$$x = \gamma_1 + \theta_1 z_1 + \theta_2 z_2 + v$$

Estimate the first-stage equation by OLS and obtain the fitted value

$$\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z_1 + \hat{\theta}_2 z_2$$

We have combined the two instruments z_1 and z_2 into the single instrument $\hat{x}$. Using $\hat{x}$ as an instrument for x leads to two sample-moment conditions,

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum\hat{x}_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

Solving these conditions, and using $\bar{\hat{x}} = \bar{x}$, we have

$$\hat{\beta}_2 = \frac{\sum\left(\hat{x}_i - \bar{\hat{x}}\right)(y_i - \bar{y})}{\sum\left(\hat{x}_i - \bar{\hat{x}}\right)(x_i - \bar{x})} = \frac{\sum(\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum(\hat{x}_i - \bar{x})(x_i - \bar{x})}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

The estimates obtained using these formulas are identical to the IV/2SLS estimates obtained by applying least squares to $y = \beta_1 + \beta_2 \hat{x} + e^*$. If we have more than two instrumental variables we apply the same strategy of combining several instruments into one.

EXAMPLE 10.4 | Using Surplus Instruments in the Simple Wage Equation

Father's education is also a potential instrument for daughter's education. Using the 428 observations in the data file *mroz*, the correlation between *FATHEREDUC* and *EDUC* is 0.4154. The first-stage equation is

$$EDUC = \gamma_1 + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

The OLS estimated first-stage equation is

$$\widehat{EDUC} = 9.4801 + 0.1564MOTHEREDUC + 0.1881FATHEREDUC$$
$$\text{(se)} \quad (0.3211) \quad (0.0358) \quad (0.0336)$$

The t-statistics for the coefficients of *MOTHEREDUC* and *FATHEREDUC* are 4.37 and 5.59, respectively, and are significant at the 1% level. The test of the joint significance of the two IV is even more important than their individual significance. The F-statistic for the null hypothesis $H_0: \theta_1 = 0, \ \theta_2 = 0$ is 55.83, which is very significant, and we can conclude that at least one of the two IV coefficients is not zero based on this joint test. The importance of the F-test is discussed in Section 10.3.9.

In the second-stage equation, we replace *EDUC* by $\widehat{EDUC}$ and apply least squares to obtain the IV/2SLS estimates

$$\widehat{\ln(WAGE)} = 0.5510 + 0.0505\widehat{EDUC}$$
$$\text{(incorrect se)} \quad (0.4257) \quad (0.0335)$$

The coefficient estimates are the correct IV estimates, but the standard errors reported are incorrect. Using proper IV software yields

$$\widehat{\ln(WAGE)} = 0.5510 + 0.0505\widehat{EDUC}$$
$$\text{(se)} \quad (0.4086) \quad (0.0322)$$

10.3.8 Instrumental Variables Estimation in the Multiple Regression Model

To implement instrumental variables estimation in a multiple regression equation, we need estimation formulas that are more general than equation (10.17). To extend our analysis to a more general setting, consider the multiple regression model $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$. Suppose that among the explanatory variables we know, or suspect, that x_K is an endogenous variable correlated with the error term. The first $K-1$ variables $\left(x_1 = 1, x_2, \ldots, x_{K-1}\right)$ are **exogenous variables** that are uncorrelated with the error term e—they are "included" instruments. Instrumental variables estimation can be carried out using a two-step process, with an OLS regression in each step.

The **first-stage regression** has the endogenous variable x_K on the left-hand side, and **all exogenous and instrumental variables** on the right-hand side. If we have L "external" instrumental variables (we are ***L***ucky to have them) that are from outside the model $z_1, z_2, \ldots, z_L$, then the first-stage regression is

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_L z_L + v_K \tag{10.20}$$

where v_K is a random error term that is uncorrelated with all the right-hand side variables. Estimate the first-stage regression (10.20) by OLS and obtain the fitted value

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \cdots + \hat{\theta}_L z_L \tag{10.21}$$

The fitted value $\hat{x}_K$ is the optimal combination of all the exogenous and instrumental variables.

The **second-stage regression** is based on the original specification with $\hat{x}_K$ replacing x_K,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K \hat{x}_K + e^* \tag{10.22}$$

where e^* is an error term. OLS estimation of (10.22) is justified because in large samples e^* is uncorrelated with the explanatory variables, including $\hat{x}_K$. The OLS estimators from this equation, $\hat{\beta}_1, \ldots, \hat{\beta}_K$, are the **instrumental variables** (IV) estimators, and, because they can be obtained by two least squares regressions, they are also popularly known as the **two-stage least squares** (2SLS) estimators. We will refer to them as IV or 2SLS or IV/2SLS estimators. In the general case with more than one endogenous variable on the right-hand side the steps are similar and are discussed in Section 10.3.10.

We can use the standard formulas for estimator variances and covariances for the least squares estimator of (10.22), which we described in Section 5.3.1, with one modification. While we can use two least squares estimations to obtain proper estimates, least squares software does not produce correct standard errors and t-values. The IV/2SLS estimator of the error variance is based on the residuals from the original model, $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$, so that the proper estimator of the error variance σ^2 is the general version of equation (10.18b)

$$\hat{\sigma}^2_{IV} = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right)^2}{N - K}$$

Econometric software will automatically use the proper variance estimator if a two-stage least squares or instrumental variables estimation option is chosen. Using the IV/2SLS estimated standard errors from (10.22), we can carry out t-tests and construct interval estimates of parameters that are valid in large samples. Furthermore, the usual tests of joint hypotheses are valid in large samples **if** the instrumental variables are not weak.

It is informative to recall the discussion in Section 6.4.1. Usually the coefficient of the endogenous variable is most interesting. Thinking about our wage equation example, the coefficient of *EDUC*, years of education, is of critical importance. Let $SSE_{\hat{x}_K}$ be the sum of squared residuals from the regression of $\hat{x}_K$ on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$, then, in large samples,

$$\hat{\beta}_K \overset{a}{\sim} N\left[\beta_K, \operatorname{var}\left(\hat{\beta}_K\right)\right]$$

and the variance estimate is

$$\widehat{\operatorname{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}} \tag{10.23}$$

Equation (10.23) shows that the variance of $\hat{\beta}_K$, the instrumental variables estimator of β_K, depends on, $SSE_{\hat{x}_K}$, the variation in $\hat{x}_K$ that is *not* explained by $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$. See equation (6.33) and the surrounding discussion. Because this is such an important concept we return to it in Section 10.3.9 when analyzing "weak" instrumental variables.

EXAMPLE 10.5 | IV/2SLS Estimation in the Wage Equation

In addition to education a worker's experience is also important in determining their wage. Because additional years of experience have a declining marginal effect on wage use the quadratic model

$$\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 EDUC + e$$

where *EXPER* is years of experience. This is the same specification as in Example 10.1. We assume that *EXPER* is an **exogenous** variable that is uncorrelated with the worker's ability and therefore uncorrelated with the random error e. Two instrumental variables for years of education, *EDUC*, are mother's and father's years of education, *MOTHEREDUC* and *FATHEREDUC*, introduced in the previous examples. The first-stage equation is

$$EDUC = \gamma_1 + \gamma_2 EXPER + \gamma_3 EXPER^2 + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

Using the 428 observations in the data file *mroz* the estimated first-stage equation is reported in Table 10.1. The IV/2SLS estimates, with correctly computed standard errors, are

$$\begin{aligned} \widehat{\ln(WAGE)} &= 0.0481 + 0.0442 EXPER \\ (\text{se}) \quad & (0.4003)\ (0.0134) \\ & - 0.0009 EXPER^2 + 0.0614 EDUC \\ & (0.0004) \qquad (0.0314) \end{aligned}$$

The estimated return to education is approximately 6.1%, and the estimated coefficient is statistically significant with a $t = 1.96$.

TABLE 10.1 First-Stage Equation

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	9.1026	0.4266	21.3396	0.0000
EXPER	0.0452	0.0403	1.1236	0.2618
$EXPER^2$	−0.0010	0.0012	−0.8386	0.4022
MOTHEREDUC	0.1576	0.0359	4.3906	0.0000
FATHEREDUC	0.1895	0.0338	5.6152	0.0000

10.3.9 Assessing Instrument Strength Using the First-Stage Model

In Section 10.3.4, we emphasized the importance of a strong instrument when estimating a simple regression model with an endogenous explanatory variable. There the assessment of the instrument's strength was based on the correlation between the endogenous variable x and the instrument z. In a multiple regression measuring instrument strength is more complicated. The first-stage regression is a key tool in assessing whether an instrument is "strong" or "weak" in the multiple regression setting.

Case 1: Assessing the Strength of One Instrumental Variable Suppose that x_K is endogenous and we have available one external instrumental variable z_1. In terms of the notation above $L = 1$. The first-stage regression equation is

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + v_K \tag{10.24}$$

In a simple regression model, we can look for instrument strength in the correlation between the endogenous variable and the instrument. In the multiple regression model, we must deal with the other exogenous variables $(x_2, \ldots, x_{K-1})$. The key to assessing the strength of the instrumental variable z_1 is the strength of its relationship to x_K **after** controlling for the effects of all the other exogenous variables. This, however, is exactly the purpose of multiple regression analysis. The coefficient θ_1 in the first-stage regression (10.24) measures the effect of z_1 on x_K after accounting for the effects of the other variables.

Not only must there be an effect of z_1 on x_K but also it must be a **statistically significant** effect. How significant? *Very significant*. To reject the hypothesis that the instrument z_1 is weak, a rule of thumb is that the F-test statistic for the null hypothesis $H_0: \theta_1 = 0$ in equation (10.24) should be greater than 10. Using the relationship between the t- and F-tests, $t^2 = F$ described in Section 6.1.3, this translates into the absolute t-statistic for significance being greater than 3.16, which is larger than the usual 5% critical values ±1.96 or the 1% critical values ±2.58. The $F > 10$ rule has been refined by econometric researchers Stock and Yogo, and we discuss their analysis in Appendix 10A. Estimates and tests based on an IV estimator are unreliable when instruments are weak.

Further Analysis of Weak Instruments[1] Another way to illustrate this point is the following. The logic may seem a bit cumbersome, but the final result will be intuitively pleasing.

[1]This section is more advanced.

In Section 10.3.8, we argued that the approximate large sample variance of the IV estimator of β_K is

$$\widehat{\text{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}}$$

where $SSE_{\hat{x}_K}$ is the sum of squared residuals from the regression of $\hat{x}_K$ on $\left(x_2, x_3, \ldots, x_{K-1}\right)$, where $\hat{x}_K$ is the fitted value from the first-stage regression (10.24),

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1$$

By taking one more step, we can obtain an insight into how important the first-stage regression results can be. Let us consider a regression of $\hat{x}_K$ on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$ and z_1. We do not need to do this in practice; we know it will result in a perfect fit, with an $R^2 = 1$. Nevertheless, let us follow the Frisch–Waugh–Lovell approach described in Section 5.2.4.

- First, partial out $\mathbf{x}_{exog}$ from $\hat{x}_K$ and obtain the residuals $\tilde{\hat{x}}_K$.
- Second, partial out $\mathbf{x}_{exog}$ from the instrument z_1 and obtain the residuals $\tilde{z}_1$. The sum of squared residuals is $\sum \tilde{z}^2_{i1}$.
- Regress $\tilde{\hat{x}}_K$ on $\tilde{z}_1$, with no constant. The estimated coefficient is $\hat{\theta}_1$, $R^2 = 1$, and the fitted value $\hat{\theta}_1 \tilde{z}_1$ exactly equals $\tilde{\hat{x}}_K$!
- Because $\tilde{\hat{x}}_K = \hat{\theta}_1 \tilde{z}_1$, we can write $SSE_{\hat{x}_K} = \sum \tilde{\hat{x}}^2_{iK} = \sum \left(\hat{\theta}_1 \tilde{z}_{i1}\right)^2 = \hat{\theta}^2_1 \sum \tilde{z}^2_{i1}$.

The result is an alternative expression for the large sample variance of the IV estimator of β_K given in (10.23),

$$\text{var}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}} = \frac{\hat{\sigma}^2_{IV}}{\hat{\theta}^2_1 \sum \tilde{z}^2_{i1}} \tag{10.25}$$

What factors contribute to the precision of the IV estimator of β_K? The first important factor is the magnitude of the estimate $\hat{\theta}_1$ from the first-stage regression. It is important that this coefficient is **large**! Second, how much variation is there in the external instrument z_1 after removing the linear effects of the included exogenous variables, $\mathbf{x}_{exog}$? What is important is the amount of variation in z_1 **not explained** by the included exogenous variables $\mathbf{x}_{exog}$. Ideally z_1 would be uncorrelated with $\mathbf{x}_{exog}$ and exhibit large variation. If $\hat{\theta}_1$ is numerically small, or if z_1 is highly correlated with $\mathbf{x}_{exog}$, or exhibits little variation, then the precision of the IV estimator $\hat{\beta}_K$ will be worse.

Case 2: Assessing the Strength of More Than One Instrumental Variable

Suppose that x_K is endogenous and we have available L external instrumental variables, z_1, z_2, …, z_L. For a single endogenous variable, we need only a single instrument. Sometimes more instruments are available, and having more strong instruments may improve the instrumental variables estimator. The first-stage regression equation is now

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \overbrace{\theta_1 z_1 + \cdots + \theta_L z_L}^{\text{external } IV} + v_K \tag{10.26}$$

What we require is that **at least one** of the instruments be strong. Given the nature of the requirement, a joint F-test of the null hypothesis $H_0 : \theta_1 = 0, \theta_2 = 0, \ldots, \theta_L = 0$ in (10.26) is relevant, because the alternative is that at least one of the θ_i coefficients is nonzero. If the F-test statistic value is sufficiently large, roughly $F > 10$, we reject the hypothesis that the instruments are "weak" and can proceed with instrumental variables estimation. If the F-value is not sufficiently large, then instrumental variables and **two-stage least squares estimation** is quite possibly worse than "ordinary" least squares.

The fitted value from the first-stage regression (10.26) is

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots + \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \cdots + \hat{\theta}_L z_L$$

Applying the Frisch–Waugh–Lovell Theorem, as in the previous section, we find that

$$\widehat{\text{var}}\left(\hat{\beta}_K\right) = \frac{\hat{\sigma}^2_{IV}}{\sum\left(\hat{\theta}_1 \tilde{z}_{i1} + \hat{\theta}_2 \tilde{z}_{i2} + \cdots + \hat{\theta}_L \tilde{z}_{iL}\right)^2} \tag{10.27}$$

where $\tilde{z}_{il}$ is the ith residual from a regression of z_l on $\mathbf{x}_{exog} = \left(x_1 = 1, x_2, x_3, \ldots, x_{K-1}\right)$. The precision of the IV estimator of β_K depends on the magnitudes of the first-stage coefficients and the unexplained components of the external instrumental variables.

EXAMPLE 10.6 | Checking Instrument Strength in the Wage Equation

In Example 10.5, there is only one potentially endogenous variable in the wage equation, *EDUC*. The minimum number of instrumental variables is one. Given two instruments, we require that at least one of them be significant in the first-stage equation. The F-test null hypothesis is that both coefficients, θ_1 and θ_2, are zero, and if we reject this null hypothesis we conclude that at least one of them is nonzero. In the first-stage regression in Table 10.1, the estimated coefficient of *MOTHEREDUC* is 0.1576 with a t-value of 4.39, and the estimated coefficient of *FATHEREDUC* is 0.1895 with a t-value of 5.62. The F-statistic value for the null hypothesis that both these coefficients are zero is 55.40, which is significant at the 1% level, but more importantly it is larger than the rule-of-thumb threshold, $F > 10$. In addition to the vitally important F-statistic, the goodness-of-fit measures R^2 and $\overline{R}^2$ are sometimes reported. For the first-stage equation in Table 10.1, these values are $R^2 = 0.1527$ and $\overline{R}^2 = 0.1467$.

Partial Correlation and Partial R^2

In addition to the first-stage F-statistic, R^2 and adjusted-R^2, a partial correlation or partial-R^2 are informative. Applying the partialling-out strategy of the Frisch–Waugh–Lovell Theorem is another way to examine instrument strength. The included exogenous variables in the wage equation are $\mathbf{x}_{exog} = \left(x_1 = 1, EXPER, EXPER^2\right)$. Regress *EDUC* on $\mathbf{x}_{exog}$ and obtain the residuals, *REDUC*.

Suppose that we are using the single instrument *MOTHEREDUC*. Regress *MOTHEREDUC* on $\mathbf{x}_{exog}$ and obtain the residuals, *RMOM*. These residual variables have the included exogenous variables partialled-out. That is, we have removed the linear influences of the included exogenous variables from the endogenous variable *EDUC* and the external IV, *MOTHEREDUC*. The correlation between *REDUC* and *RMOM* is called a **partial correlation**, and in this case it is 0.3854. The R^2 from a regression of *REDUC* on *RMOM* is called the partial-R^2, and in this case it is 0.1485. Because we have one endogenous variable and one external IV, the partial-$R^2 = 0.1485$ is the square of the partial correlation, $0.3854^2 = 0.1485$.

If there are more external instruments, the partial-R^2 is the R^2 of the partialled-out endogenous variable on all the partialled-out external IV. Add *FATHEREDUC* as an IV, regress it on $\mathbf{x}_{exog}$ and obtain the residuals, *RDAD*. The partial-R^2 is then the R^2 from the regression of *REDUC* on *RMOM* and *RDAD*. In this case, partial-$R^2 = 0.2076$ and the adjusted partial-$R^2 = 0.2038$.

10.3.10 Instrumental Variables Estimation in a General Model

To extend our analysis to a more general setting, consider the multiple regression model $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$. Suppose that among the explanatory variables $\left(x_1 = 1, x_2, \ldots, x_K\right)$ we know, or suspect, that several may be correlated with the error term e. Divide the variables into two groups, with the first G variables $\left(x_1 = 1, x_2, \ldots, x_G\right)$ being exogenous variables that are uncorrelated with the error term e. The second group of $B = K - G$ variables $\left(x_{G+1}, x_{G+2}, \ldots, x_K\right)$

is correlated with the regression error, and thus they are endogenous. The multiple regression model, including all K variables, is then

$$y = \overbrace{\beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G}^{G \text{ exogenous variables}} + \overbrace{\beta_{G+1} x_{G+1} + \cdots + \beta_K x_K}^{B \text{ endogenous variables}} + e \tag{10.28}$$

In order to carry out IV estimation we must have at least as many instrumental variables as we have **endogenous variables**. Suppose we have L external instrumental variables, from outside the model, $z_1, z_2, \ldots, z_L$. Such notation is invariably confusing and cumbersome. It may help to keep things straight to think of $G = Good$ explanatory variables and $B = Bad$ explanatory variables and $L = Lucky$ instrumental variables, since we are lucky to have them. Then we have *The Good, the Bad, and the Lucky*.

It is a necessary condition for IV estimation that $L \geq B$. If $L = B$ then there are just enough instrumental variables to carry out IV estimation. The model parameters are said to be **just-identified** or **exactly identified** in this case. The term **identified** is used to indicate that the model parameters can be consistently estimated. If $L > B$ then we have more instruments than are necessary for IV estimation, and the model is said to be **overidentified**.

To implement IV/2SLS, estimate B first-stage equations, one for each explanatory variable that is endogenous. On the left-hand side of the first-stage equations, we have an endogenous variable. On the right-hand side, we have *all* the exogenous variables, including the G explanatory variables that are exogenous, *and* the L instrumental variables, which also must be exogenous. The B first-stage equations are

$$x_{G+j} = \gamma_{1j} + \gamma_{2j} x_2 + \cdots + \gamma_{Gj} x_G + \theta_{1j} z_1 + \cdots + \theta_{Lj} z_L + v_j, \quad j = 1, \ldots, B \tag{10.29}$$

The first-stage parameters (γ's and θ's) take different values in each equation, which is why they have a "j" subscript. We have omitted the observation subscript for simplicity. Since the right-hand side variables are all exogenous, we can estimate (10.29) by OLS. Then obtain the fitted values

$$\hat{x}_{G+j} = \hat{\gamma}_{1j} + \hat{\gamma}_{2j} x_2 + \cdots + \hat{\gamma}_{Gj} x_G + \hat{\theta}_{1j} z_1 + \cdots + \hat{\theta}_{Lj} z_L, \quad j = 1, \ldots, B$$

This comprises the first stage of two-stage OLS estimation.

In the second stage of estimation we apply least squares to

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1} \hat{x}_{G+1} + \cdots + \beta_K \hat{x}_K + e^* \tag{10.30}$$

This two-stage estimation process leads to proper instrumental variables estimates, but it should not be done this way in applied work. Use econometric software designed for two-stage least squares or instrumental variables estimation so that standard errors, t-statistics, and other test statistics will be computed properly.

Assessing Instrument Strength in a General Model

The F-test for **weak instruments** discussed in Section 10.3.9 is not valid for models having more than one endogenous variable on the right side of the equation. Consider the model in (10.28) with $B = 2$,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G + \beta_{G+1} x_{G+1} + \beta_{G+2} x_{G+2} + e \tag{10.31}$$

where $x_2, \ldots, x_G$ are exogenous and uncorrelated with the error term e, while x_{G+1} and x_{G+2} are endogenous. Suppose that we have two external instrumental variables z_1 and z_2, with z_1 being a good instrument for both x_{G+1} and x_{G+2}. The weak instrument F-test may be significant in each first-stage equation even if z_2 is an irrelevant instrument and not at all related to x_{G+1} or x_{G+2}. In such a case, we might conclude that we have two valid instruments when we have only one.

The first-stage equations in this case are

$$x_{G+1} = \gamma_{11} + \gamma_{21}x_2 + \cdots + \gamma_{G1}x_G + \theta_{11}z_1 + \theta_{21}z_2 + v_1$$

$$x_{G+2} = \gamma_{12} + \gamma_{22}x_2 + \cdots + \gamma_{G2}x_G + \theta_{12}z_1 + \theta_{22}z_2 + v_2$$

The weak instrument F-test in the first equation is for the joint significance of θ_{11} and θ_{21}, $H_0: \theta_{11} = 0$, $\theta_{21} = 0$, with the alternative hypothesis that at least *one* of these coefficients is not zero. If θ_{11} is statistically significant, then the joint null hypothesis may be rejected even if $\theta_{21} = 0$. Similarly in the second equation we can obtain a significant F-test outcome even if z_2 is irrelevant as an instrument for x_{G+1} as long as z_1 is statistically significant. In this case we have two individually significant F-tests despite the fact that only one valid instrument z_1 is available, and thus the model in (10.31) is not identified. The more general test required for this case, which builds on the concept of "partial correlation" is discussed in Appendix 10A.

10.3.11 Additional Issues When Using IV Estimation

In this section, we discuss some issues related to IV estimation.

Hypothesis Testing with Instrumental Variables Estimates We may be interested in testing hypotheses about the regression parameters based on the two-stage least squares/instrumental variables estimates. When testing the null hypothesis $H_0: \beta_k = c$, use of the test statistic $t = \left(\hat{\beta}_k - c\right) \Big/ \text{se}\left(\hat{\beta}_k\right)$ is valid in large samples. We know that as $N \to \infty$, the $t_{(N-K)}$ distribution converges to the standard normal distribution $N(0, 1)$. If the degrees of freedom $N - K$ are large, then critical values from the two distributions will be very close. It is common, but not universal, practice to use critical values, and p-values, based on the $t_{(N-K)}$ distribution rather than the more strictly appropriate $N(0, 1)$ distribution. The reason is that tests based on the t-distribution tend to work better in samples of data that are not large.

Another issue is whether to use standard errors that are "robust" to the presence of heteroskedasticity (in cross-section data) or autocorrelation and heteroskedasticity (in time-series data). These options were described in Chapters 8 and 9 for the linear regression model, and they are also available in most software packages for IV estimation. Such corrections to standard errors require large samples in order to work properly.

When using software to test a joint hypothesis, such as $H_0: \beta_2 = c_2, \beta_3 = c_3$, the test may be based on the chi-square distribution with the number of degrees of freedom equal to the number of hypotheses (J) being tested. The test itself may be called a Wald test, or a likelihood ratio (LR) test, or a Lagrange multiplier (LM) test. These testing procedures are all asymptotically equivalent and are discussed in Appendix C.8.4. However, the test statistic reported may also be called an F-statistic with J numerator degrees of freedom and $N - K$ denominator degrees of freedom. This F-value is often calculated by dividing one of the chi-square tests statistics, such as the Wald statistic, by J. The motivation for using the F-test is to achieve better performance in small samples. Asymptotically, the tests will all lead to the same conclusion. See Chapter 6, Appendix 6A, for some related discussion. Once again, joint tests can be made "robust" to potential heteroskedasticity or autocorrelation problems, and this is an option with many software packages.

Generalized Method-of-Moments Estimation If heteroskedasticity or serial correlation is present in a model with one or more endogenous variables, then using instrumental variables estimation with a "robust" covariance matrix ensures that interval estimators, hypothesis tests and prediction intervals use a valid standard error. However, using an instrumental

variables estimator with a robust covariance matrix estimator does not *improve* the efficiency of the estimator, just like using the OLS estimator with a robust covariance matrix estimator does not improve its efficiency. In Chapters 8 and 9 we introduced a **generalized least squares estimator** for linear regression models with error terms that are heteroskedastic and/or serially correlated. In the same way, there is a **generalized method-of-moments (GMM) estimator** that is "asymptotically" more efficient than the instrumental variables estimator. Being "asymptotically more efficient" means that the GMM estimator has smaller variances than the IV estimator in large samples. In order to obtain the gain, we must have at least one surplus instrument. The gain in efficiency is obtained by building into the estimator a heteroskedasticity and/or serial correlation correction. Despite the fact that the GMM estimator improves the large sample precision of estimation its actual performance in samples that are not large might not be good. And like the IV estimator, good instruments are required. Theoretically, the GMM estimator is very attractive because it is a general estimation approach that includes the OLS estimator, the GLS estimator and IV/2SLS as special cases.

The GMM estimation procedure is built into econometric software packages but their proper usage requires an in-depth study of the methodology, which is beyond the scope of this book. It is one of the few topics that is difficult to explain without the tools of matrix algebra. Advanced readers can consult William Greene (2018) *Econometric Analysis, Eighth Edition*, Pearson Prentice-Hall, Chapter 13.

Goodness-of-Fit with Instrumental Variables Estimates We discourage the use of measures like R^2 outside the context of OLS estimation. When there are endogenous variables on the right-hand side of a regression equation, the concept of measuring how well the variation in y is explained by the x variables breaks down, because as we discussed in Section 10.2, these models exhibit feedback. This logical problem is paired with a numerical one. If our model is $y = \beta_1 + \beta_2 x + e$, then the IV residuals are $\hat{e} = y - \hat{\beta}_1 - \hat{\beta}_2 x$. Many software packages will report the goodness-of-fit measure $R^2 = 1 - \sum \hat{e}_i^2 / \sum (y_i - \bar{y})^2$. Unfortunately, this quantity can be negative when based on IV estimates.

10.4 Specification Tests

We have shown that if an explanatory variable is correlated with the regression error term, the OLS estimator fails. If a strong instrumental variable is available, the IV estimator is consistent and approximately normally distributed in large samples. But if we use a weak instrument, or an instrument that is invalid in the sense that it is not uncorrelated with the regression error, then IV estimation can be as bad as, or worse than, using the OLS estimator. We addressed how to detect weak instruments in Section 10.3.9, and go into much greater detail on this problem in Appendix 10A. In this section we ask two other important questions that must be answered in each situation in which instrumental variables estimation is considered:

1. Can we test for whether x is correlated with the error term? This might give us a guide for when to use least squares and when to use IV estimators.
2. Can we test if our instrument is valid, and uncorrelated with the regression error, as required?

10.4.1 The Hausman Test for Endogeneity

In the previous sections, we discussed the fact that the least squares estimator fails if there is correlation between an explanatory variable and the error term. We also provided an estimator, the instrumental variables estimator, that can be used when the least squares estimator fails.

The question we address in this section is how to test for the presence of a correlation between an explanatory variable and the error term, so that we can use the appropriate estimation procedure.

The null hypothesis is $H_0: \text{cov}(x_i, e_i) = 0$ against the alternative that $H_1: \text{cov}(x_i, e_i) \neq 0$. The idea of the test is to compare the performance of the OLS estimator to an instrumental variables estimator. Under the null and alternative hypotheses, we know the following:

- If the null hypothesis is true, both the OLS estimator b and the instrumental variables estimator $\hat{\beta}$ are consistent. Thus, in large samples the difference between them converges to zero. That is, $q = (b - \hat{\beta}) \to 0$. Naturally, if the null hypothesis is true, use the more efficient estimator, which is the least squares estimator.
- If the null hypothesis is false, the OLS estimator is not consistent, and the instrumental variables estimator is consistent. Consequently, the difference between them does not converge to zero in large samples. That is, $q = (b - \hat{\beta}) \to c \neq 0$. If the null hypothesis is not true, use the instrumental variables estimator, which is consistent.

There are several forms of the test, usually called the **Hausman test** in recognition of econometrician Jerry Hausman's pioneering work on this problem, for these null and alternative hypotheses. One form of the test directly examines the differences between the least squares and instrumental variables estimates, as we have described above. Some computer software programs implement this test for the user, which can be computationally difficult to carry out.[2]

An alternative form of the test is very easy to implement, and is the one we recommend. See Section 10.4.2 for an explanation of the test's logic. In the regression $y_i = \beta_1 + \beta_2 x_i + e_i$, we wish to know whether x_i is correlated with e_i. Let z_1 and z_2 be instrumental variables for x. At minimum, one instrument is required for each variable that might be correlated with the error term. Then carry out the following steps:

1. Estimate the first-stage model $x = \gamma_1 + \theta_1 z_1 + \theta_2 z_2 + v$ by OLS, including on the right-hand side all instrumental variables and all exogenous variables not suspected to be endogenous, and obtain the residuals
$$\hat{v} = x - \hat{\gamma}_1 - \hat{\theta}_1 z_1 - \hat{\theta}_2 z_2$$
If more than one explanatory variable is being tested for endogeneity, repeat this estimation for each one.
2. Include the residuals computed in step 1 as an explanatory variable in the original regression, $y = \beta_1 + \beta_2 x + \delta \hat{v} + e$. Estimate this "artificial regression" by OLS, and employ the usual t-test for the hypothesis of significance:
$$H_0: \delta = 0 \quad (\text{no correlation between } x_i \text{ and } e_i)$$
$$H_1: \delta \neq 0 \quad (\text{correlation between } x_i \text{ and } e_i)$$
3. If more than one variable is being tested for endogeneity, the test will be an F-test of joint significance of the coefficients on the included residuals.

The t- and F-tests in steps two and three can be made robust if heteroskedasticity and/or autocorrelation are potential problems.

[2] Some software packages compute Hausman tests with K, or $K - 1$, degrees of freedom, where K is the total number of regression parameters. This is incorrect. Use the correct degrees of freedom B, equal to the number of potentially endogenous right-hand-side variables (see 10.28).

10.4.2 The Logic of the Hausman Test[3]

In Section 10.4.1, we presented the Hausman test for whether or not an explanatory variable is endogenous using an artificial regression. Let us explore how this test works. The simple regression model is

$$y = \beta_1 + \beta_2 x + e \tag{10.32}$$

If x is correlated with the error term e, then x is endogenous and the OLS estimator is biased and inconsistent.

An instrumental variable z must be correlated with x but uncorrelated with e in order to be valid. A correlation between z and x implies that there is a linear association between them. This means that we can describe their relationship as a regression

$$x = \gamma_1 + \theta_1 z + v \tag{10.33}$$

This is the first-stage equation introduced in Section 10.3.6. It is a predictive model with the base assumption $E(x|z) = \gamma_1 + \theta_1 z$. The conditional mean of the endogenous variable x is linearly related to the instrumental variable z. The error term v is simply $v = x - (\gamma_1 + \theta_1 z)$ so that the two sides of (10.33) are equal. There is a correlation between x and z if, and only if, $\theta_1 \neq 0$. We can divide x into two parts, a systematic part and a random part, as

$$x = E(x|z) + v \tag{10.34}$$

where $E(x|z) = \gamma_1 + \theta_1 z$. If we knew γ_1 and θ_1, we could substitute (10.34) into the simple regression model (10.32) to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2 \left[E(x|z) + v\right] + e \\ &= \beta_1 + \beta_2 E(x|z) + \beta_2 v + e \end{aligned} \tag{10.35}$$

Now, suppose for a moment that $E(x|z)$ and v can be observed and are viewed as explanatory variables in the regression $y = \beta_1 + \beta_2 E(x|z) + \beta_2 v + e$. Will least squares work when applied to this equation? The explanatory variable $E(x|z)$ depends only on z and it is not correlated with the error term e if z is a valid instrument. The endogeneity problem, if there is one, comes from a correlation between v (the random part of x) and e. In fact, in the regression (10.32) any correlation between x and e implies correlation between v and e because $v = x - E(x|z)$.

We cannot exactly create the partition in (10.34) because we do not know γ_1 and θ_1. However, we can consistently estimate the first-stage equation (10.33) by OLS. Doing so, we obtain the fitted first-stage equation $\hat{x} = \widehat{E(x|z)} = \hat{\gamma}_1 + \hat{\theta}_1 z$ and the residuals $\hat{v} = x - \hat{x}$. Rearrange these to obtain an estimated analog of (10.34),

$$x = E(x|z) + \hat{v} = \hat{x} + \hat{v} \tag{10.36}$$

Substitute (10.36) into the original equation (10.32) to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2 \left[\hat{x} + \hat{v}\right] + e \\ &= \beta_1 + \beta_2 \hat{x} + \beta_2 \hat{v} + e \end{aligned} \tag{10.37}$$

To reduce confusion, and avoid β_2 appearing twice in same equation, let the coefficient of $\hat{v}$ be denoted as γ, so that (10.37) becomes

$$y = \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e \tag{10.38}$$

[3]Contains advanced material.

If we omit $\hat{v}$ from (10.38) the regression becomes

$$y = \beta_1 + \beta_2 \hat{x} + e \tag{10.39}$$

The least squares estimates of β_1 and β_2 in (10.39) *are* the IV/2SLS estimates discussed in Section 10.3.6. Then, recall from Section 6.6.1, equation (6.23), that if we omit a variable from a regression that is uncorrelated with the included variable(s) there is no omitted variables bias, and in fact the least squares estimates are unchanged! This holds true in (10.39) because the least squares residuals $\hat{v}$ are uncorrelated with $\hat{x}$ and the intercept variable. Thus, the least squares estimates of β_1 and β_2 in (10.38) and (10.39) are identical and are equal to the IV/2SLS estimates. Consequently, the least squares estimators of β_1 and β_2 in (10.38) are consistent whether or not x is exogenous, because they are the IV estimators.

What about γ? If x is exogenous, and hence v and e are uncorrelated, then the least squares estimator of γ in (10.38) will also converge in large samples to β_2. However, if x is endogenous then the least squares estimator of γ in (10.38) will *not* converge to β_2 in large samples because $\hat{v}$, like v, is correlated with the error term e. This observation makes it possible to test for whether x is exogenous by testing the equality of the estimates of β_2 and γ in (10.38). If we reject the null hypothesis $H_0: \beta_2 = \gamma$ then we reject the exogeneity of x, and conclude that it is endogenous.

Carrying out the test is made simpler by playing a trick on (10.38). Add and subtract $\beta_2 \hat{v}$ to the right-hand side to obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e + \beta_2 \hat{v} - \beta_2 \hat{v} \\ &= \beta_1 + \beta_2 (\hat{x} + \hat{v}) + (\gamma - \beta_2) \hat{v} + e \\ &= \beta_1 + \beta_2 x + \delta \hat{v} + e \end{aligned} \tag{10.40}$$

Thus, instead of testing $H_0: \beta_2 = \gamma$ we can simply use an ordinary t-test of the null hypothesis $H_0: \delta = 0$ in (10.40), which is exactly the test we described in Section 10.4.1. This is much nicer because software automatically prints out the t-statistic for this hypothesis test. This test can be made robust to heteroskedasticity and/or autocorrelation if desired.

10.4.3 Testing Instrument Validity

A valid instrument z must be contemporaneously uncorrelated with the regression error term, so that $\text{cov}(z_i, e_i) = 0$. If this condition fails then the resulting moment condition, like (10.16), is invalid and the IV estimator will not be consistent. Unfortunately, not every instrument can be tested for validity. In order to compute the IV estimator for an equation with B possibly endogenous variables, we must have at least B instruments. The validity of this minimum number of required instruments cannot be tested. In the case in which we have $L > B$ instruments available, we can test the validity of the $L - B$ extra, or surplus, moment conditions.[4]

An intuitive approach is the following. From the set of L instruments, form groups of B instruments and compute the IV estimates using each different group. If all the instruments are valid, then we would expect all the IV estimates to be similar. Rather than do this, there is a test of the validity of the **surplus moment conditions** that is easier to compute. The steps are

1. Compute the IV estimates $\hat{\beta}_k$ using all available instruments, including the G variables $x_1 = 1, x_2, \ldots, x_G$ that are presumed to be exogenous, and the L instruments $z_1, \ldots, z_L$.
2. Obtain the residuals $\hat{e}_{IV} = y - \hat{\beta}_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_K x_K$.

[4]Econometric jargon for surplus moment conditions is "overidentifying restrictions." A surplus of moment conditions means we have more than enough for identification, hence "overidentifying." Moment conditions like (10.16) can be thought of as restrictions on parameters.

3. Regress $\hat{e}_{IV}$ on all the available instruments described in step one.
4. Compute NR^2 from this regression, where N is the sample size and R^2 is the usual goodness-of-fit measure.
5. If all of the surplus moment conditions are valid, then $NR^2 \sim \chi^2_{(L-B)}$.[5] If the value of the test statistic exceeds the $100(1-\alpha)$th percentile (i.e., the critical value) from the $\chi^2_{(L-B)}$ distribution, then we conclude that at least one of the surplus moment conditions is not valid.

If we reject the null hypothesis that all the surplus moment conditions are valid, then we are faced with trying to determine which instrument(s) are invalid, and how to weed them out.

EXAMPLE 10.7 | Specification Tests for the Wage Equation

In Section 10.3.6, we examined a ln(*WAGE*) equation for married women, using the two instruments "mother's education" and "father's education" for the potentially endogenous explanatory variable education (*EDUC*).

To implement the Hausman test we first obtain the first-stage regression estimates, which are shown in Table 10.1. Using these estimates we calculate the least squares residuals $\hat{v} = EDUC - \widehat{EDUC}$. Insert the residuals in the ln(*WAGE*) equation as an extra variable, and estimate the resulting augmented regression using OLS. The resulting estimates are shown in Table 10.2.

The Hausman test of the endogeneity is based on the t-test of significance of the first-stage regression residuals, $\hat{v}$. If we reject the null hypothesis that the coefficient is zero, we conclude that education is endogenous. Note that the coefficient of the first-stage regression residuals (*VHAT*) is significant at the 10% level of significance using a two-tail test. While this is not strong evidence of the endogeneity of education, it is sufficient cause for concern to consider using instrumental variables estimation. Second, note that the coefficient estimates of the remaining variables, but not their standard errors, are identical to their instrumental variables estimates. This feature of the regression-based Hausman test is explained in Section 10.4.2.

In order to be valid, the instruments *MOTHEREDUC* and *FATHEREDUC* should be uncorrelated with the regression error term. As discussed in Section 10.4.3, we cannot test the validity of both instruments, only the "overidentifying" or surplus instrument. Since we have two instruments and only one potentially endogenous variable, we have $L - B = 1$ extra instrument. The test is carried out by regressing the residuals from the ln(*WAGE*) equation, calculated using the instrumental variables estimates, on all available exogenous and instrumental variables. The test statistic is NR^2 from this artificial regression, and R^2 is the usual goodness-of-fit measure. If the surplus instruments are valid, then the test statistic has an asymptotic $\chi^2_{(1)}$ distribution, where the degrees of freedom are the number of surplus instruments. If the test statistic value is greater than the critical value from this distribution, then we reject the null hypothesis that the surplus instrument is valid. For the artificial regression $R^2 = 0.000883$, and the test statistic value is $NR^2 = 428 \times 0.000883 = 0.3779$. The 0.05 critical value for the chi-square distribution with one degree of freedom is 3.84, so we fail to reject the surplus instrument as valid. With this result we are reassured that our instrumental variables estimator for the wage equation is consistent.

TABLE 10.2 Hausman Test Auxiliary Regression

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.0481	0.3946	0.1219	0.9030
EDUC	0.0614	0.0310	1.9815	0.0482
EXPER	0.0442	0.0132	3.3363	0.0009
$EXPER^2$	−0.0009	0.0004	−2.2706	0.0237
VHAT	0.0582	0.0348	1.6711	0.0954

[5] This test is valid if errors are homoskedastic and is sometimes called the Sargan test. If the errors are heteroskedastic, there is a more general test called Hansen's J-test that is provided by some software. A very advanced reference is Hayashi, *Econometrics*, Princeton, 2000, pp. 227–228.

10.5 Exercises

10.5.1 Problems

10.1 Using state level data, a researcher wishes to examine the relationship between the median rent paid (*RENT*) as a function of median house values (*MDHOUSE* in $1000). The percentage of the state population living in an urban area (*PCTURBAN*) is used as an additional control. Use the results in Table 10.3 to answer the following questions.

TABLE 10.3 Estimates for Exercise 10.1

	(1) *RENT*	(2) *MDHOUSE*	(3) *MDHOUSE*	(4) *RENT*	(5) *RENT*	(6) *EHAT*
C	125.9	−19.78	7.225	121.1	121.1	−53.50
	(14.19)	(10.23)	(8.936)	(12.87)	(15.51)	(22.66)
PCTURBAN	0.525	0.205	0.616	0.116	0.116	−0.257
	(0.249)	(0.113)	(0.131)	(0.254)	(0.306)	(0.251)
MDHOUSE	1.521			2.184	2.184	
	(0.228)			(0.282)	(0.340)	
FAMINC		2.584				3.851
		(0.628)				(1.393)
REG4		15.89				−16.87
		(3.157)				(6.998)
VHAT				−1.414		
				(0.411)		
N	50	50	50	50	50	50
R^2	0.669	0.679	0.317	0.737	0.609	0.198
SSE	20259.6	3907.4	8322.2	16117.6	23925.6	19195.8

Standard errors in parentheses.

a. The OLS estimates of the model are in column (1). Why might we be concerned that *MDHOUSE*, the median price of houses, is endogenous in this regression?

b. Two instruments are considered: median family income (*FAMINC* in $1000) and a regional dummy variable *REG4*. Using the models in columns (2) and (3), test if the instruments are weak.

c. In column (4), the least squares residuals (*VHAT*) from the regression in column (2) are added as a regressor to the basic regression. The estimates are obtained using OLS. What is the usefulness of this regression? What does it indicate about the results in (1)?

d. In column (5) are IV/2SLS estimates using the instruments listed in part (b). What differences do you observe between these results and the OLS results in column (1)? Note that the estimates (though not the standard errors) are the same in columns (4) and (5). Is this a mistake? Explain.

e. In column (6) the residuals from the estimation in column (5) are regressed upon the variables shown. What information is contained in these results?

10.2 The labor supply of married women has been a subject of a great deal of economic research. Consider the following supply equation specification

$$HOURS = \beta_1 + \beta_2 WAGE + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 NWIFEINC + e$$

where *HOURS* is the supply of labor, *WAGE* is hourly wage, *EDUC* is years of education, *KIDSL6* is the number of children in the household who are less than 6 years old, and *NWIFEINC* is household income from sources other than the wife's employment.

a. Discuss the signs you expect for each of the coefficients.
b. Explain why this supply equation cannot be consistently estimated by OLS regression.
c. Suppose we consider the woman's labor market experience *EXPER* and its square, $EXPER^2$, to be instruments for *WAGE*. Explain how these variables satisfy the logic of instrumental variables.
d. Is the supply equation identified? Explain.
e. Describe the steps [not a computer command] you would take to obtain IV/2SLS estimates.

10.3 In the regression model $y = \beta_1 + \beta_2 x + e$, assume x is endogenous and that z is a valid instrument. In Section 10.3.5, we saw that $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$.

a. Divide the denominator of $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$ by $\text{var}(z)$. Show that $\text{cov}(z, x)/\text{var}(z)$ is the coefficient of the simple regression with dependent variable x and explanatory variable z, $x = \gamma_1 + \theta_1 z + v$. [*Hint*: See Section 10.2.1.] Note that this is the first-stage equation in two-stage least squares.
b. Divide the numerator of $\beta_2 = \text{cov}(z, y)/\text{cov}(z, x)$ by $\text{var}(z)$. Show that $\text{cov}(z, y)/\text{var}(z)$ is the coefficient of a simple regression with dependent variable y and explanatory variable z, $y = \pi_0 + \pi_1 z + u$. [*Hint*: See Section 10.2.1.]
c. In the model $y = \beta_1 + \beta_2 x + e$, substitute for x using $x = \gamma_1 + \theta_1 z + v$ and simplify to obtain $y = \pi_0 + \pi_1 z + u$. What are π_0, π_1, and u in terms of the regression model parameters and error and the first-stage parameters and error? The regression you have obtained is a **reduced-form** equation.
d. Show that $\beta_2 = \pi_1/\theta_1$.
e. If $\hat{\pi}_1$ and $\hat{\theta}_1$ are the OLS estimators of π_1 and θ_1, show that $\hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1$ is a consistent estimator of $\beta_2 = \pi_1/\theta_1$. The estimator $\hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1$ is an **indirect least squares** estimator.

10.4 Suppose that x is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero; it is an indicator variable. Make the assumption $E(e_i|z_i) = 0$.

a. Show that $E(y_i|z_i) = \beta_1 + \beta_2 E(x_i|z_i)$.
b. Assume $E(x_i|z_i) \neq 0$. Does z_i satisfy conditions IV1–IV3? Explain.
c. Write out the **conditional expectation** in (a) for the two cases with $z_i = 1$ and $z_i = 0$. Solve the two resulting equations for β_2.
d. Suppose we have a random sample (y_i, x_i, z_i), $i = 1, \ldots, N$. Give an intuitive argument that a consistent estimator of $E(y_i|z_i = 1)$ is the sample average of the y_i values for the subset of observations for which $z_i = 1$, which we might call $\bar{y}_1$.
e. Following the strategy in part (d) form $\bar{y}_1$, $\bar{y}_0$, $\bar{x}_1$, and $\bar{x}_0$. Show that the empirical implementation of the expression in (c) is $\hat{\beta}_{WALD} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$, which is the **Wald Estimator**, in honor of Abraham Wald.
f. Explain how $E(x_i|z_i = 1) - E(x_i|z_i = 0)$ might be viewed as a measure of the strength of the instrumental variable z_i.

10.5 Suppose that x_i is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero with probabilities p and $1 - p$, respectively, that is, $\Pr(z_i = 1) = p$ and $\Pr(z_i = 0) = 1 - p$.

a. Show that $E(z_i) = p$.
b. Show that $E(y_i z_i) = pE(y_i|z_i = 1)$.
c. Use the law of iterated expectations to show that $E(y_i) = pE(y_i|z_i = 1) + (1 - p)\,E(y_i|z_i = 0)$.
d. Substitute (a), (b), and (c) results into $E(y_i z_i) - E(y_i)E(z_i)$ to show that $\text{cov}(y_i, z_i) = p(1 - p)\,E(y_i|z_i = 1) - p(1 - p)\,E(y_i|z_i = 0)$.
e. Use the arguments in (a)–(d) to show that $\text{cov}(x_i, z_i) = p(1 - p)\left[E(x_i|z_i = 1) - E(x_i|z_i = 0)\right]$.
f. Assuming $E(e_i) = 0$ show $\left[y_i - E(y_i)\right] = \beta_2\left[x_i - E(x_i)\right] + e_i$.
g. Multiply both sides of the expression in (f) by $z_i - E(z_i)$ and take expectations to show $\text{cov}(y_i, z_i) = \beta_2 \text{cov}(x_i, z_i)$ if $\text{cov}(e_i, z_i) = 0$.
h. Using (d), (f), and (g) show that $\beta_2 = \dfrac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$
i. Show that the empirical implementation of (h) leads to $\hat{\beta}_{WALD} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$.

10.6 Suppose that x_i is endogenous in the regression $y_i = \beta_1 + \beta_2 x_i + e_i$. Suppose that z_i is an instrumental variable that takes two values, one and zero.

a. Let $N_1 = \sum z_i$ be the number of z_i values such that $z_i = 1$. Show that $\sum z_i x_i = N_1 \bar{x}_1$ where $\bar{x}_1$ is the sample average of the x_i values corresponding to $z_i = 1$.

b. Let $N_0 = N - \sum z_i = N - N_1$ be the number of z_i values such that $z_i = 0$. Show that $\sum x_i = N_1 \bar{x}_1 + N_0 \bar{x}_0$ where $\bar{x}_0$ is the sample average of the x_i values corresponding to $z_i = 0$.

c. Show that $N\sum x_i z_i - \sum z_i \sum x_i = N_1 N_0 (\bar{x}_1 - \bar{x}_0)$

d. Show that $N\sum y_i z_i - \sum z_i \sum y_i = N_1 N_0 (\bar{y}_1 - \bar{y}_0)$

e. Use the results in (c) and (d) to show that the IV estimator of β_2 in (10.17) reduces to $\hat{\beta}_2 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$.

f. The estimated variance of the IV estimator is given in (10.18a). Show that $\sum (z_i - \bar{z})(x_i - \bar{x}) = \sum z_i x_i - N\bar{z}\bar{x} = N_1 N_0 (\bar{x}_1 - \bar{x}_0)$.

g. Using the result in part (f), suppose $(\bar{x}_1 - \bar{x}_0) \simeq 0$. How does this indicate that the IV z_i is weak?

h. $\sum (z_i - \bar{z})(x_i - \bar{x}) / \sum (z_i - \bar{z})^2$ is the OLS estimate of the slope coefficient from a regression of x_i on z_i. True or False? How does this value relate to the weak instrument discussion in part (g)? If this coefficient is small, with a low t-value, does it imply that z_i is a weak IV? Explain.

10.7 Angrist and Krueger (1991) use quarter of birth as an instrumental variable to estimate the returns to schooling, using a sample of 327,509 from the 1980 census. The model of interest is $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$.

a. Let $\overline{\ln(WAGE)}$ denote the average of the natural log of weekly wage. For men born in the first quarter of the year the average is 5.8916, and for men born in the fourth quarter of the year the average is 5.9027. What is the approximate percentage difference in wages for the two groups of men?

b. The standard error of the difference in means from part (a) is 0.00274. Is the difference in $\overline{\ln(WAGE)}$ statistically significant? What is the two-tail p-value?

c. Let $\overline{EDUC}$ denote the average years of schooling. For men born in the first quarter of the year the average is 12.6881, and for men born in the fourth quarter of the year the average is 12.7969. What is the approximate percentage difference in years of schooling for the two groups of men? Is there a reason why men born in the fourth quarter have higher average schooling than men born in the first quarter?

d. The standard error of the difference in means from part (c) is 0.0132. Is the difference in $\overline{EDUC}$ statistically significant? What is the two-tail p-value.

e. Compute the Wald estimate of the return to schooling, $\hat{\beta}_{2,WALD}$ using the results above. What is the instrumental variable z being used in this case? The Wald estimator is introduced in Exercise 10.4.

f. Explain why the result in (d) is important to the success of the Wald estimator.

10.8 Knowledge is Power Program (KIPP) Schools are charter schools with largely minority students. These schools differ in a number of ways from public schools, but emphasize longer days and more time spent in school. The question is: "How much benefit is there to attending a KIPP school?"[6]

a. Let $y_i = MATH_i$ be the outcome of a math achievement test. This outcome is standardized by subtracting the average and dividing by the standard deviation, so that $y = 0$ is the average score, and $y = 1$ is a score that is one standard deviation above average, and so on. Let $x_i = ATTENDED_i$ be an indicator variable with the value one if a student attended a KIPP school and zero otherwise. In the regression $y_i = \beta_1 + \beta_2 x_i + e_i$, suppose that the OLS estimate of β_2 is $b_2 = 0.467$, with a standard error of 0.103. Based on this regression result, does attending a KIPP school seem to improve math test score? Is the estimate of the amount of improvement a meaningful amount? If the average math score of those attending the KIPP school is 0.095, what is the average score of those who do not attend the KIPP school?

b. Explain why we might worry that *ATTENDED* is an endogenous variable.

c. Offers of admission are randomly assigned to the pool of KIPP applicants. Some of those offered admission wind up attending and some do not. Let *WINNER* be an indicator variable taking the value one if a student receives an offer to attend, and zero otherwise. Suppose that 78.7% of offers to attend are accepted. Does *WINNER* satisfy the conditions for an instrumental variable?

[6] This exercise is adapted from Angrist and Pischke (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.

d. Suppose that $z_i = WINNER_i$. In the terms of this example, explain the components of

$$\beta_2 = \frac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$$

See Exercises 10.4 and 10.5 for background discussion of the expression.

e. The average math score of those receiving an offer to attend the KIPP school was −0.003, which is very close to average. The average score of those not offered a seat was −0.358, which is about one-third of a standard deviation below average. Interestingly, some students wind up attending the KIPP school despite not being randomly selected from the applicants. Assume that the proportion of students attending the KIPP school who were not "winners" be 4.6%. Obtain the Wald estimator of β_2 by replacing the population averages in part (d) with sample averages. How does this estimate compare to the OLS estimate in part (a)? Does attending a KIPP school appear to have a meaningful positive effect on scores of those attending?

10.9 Consider the wage equation used in Example 10.5. Suppose we have a variable designed to measure *ABILITY*. This variable is an index created using 10 different tests of cognitive ability. Using data on 2,178 white males in 1980, the ability variable has a sample mean of 0.04 and a standard deviation of 0.96.

a. The estimated relationship between years of education and the ability measure is $\widehat{EDUC} = 12.30 + 0.977ABILITY$ with a t-value of 25.81. Is this result consistent with the usual "omitted variables bias" explanation of the endogeneity of education? Explain.

b. Using these data and the model in Example 10.5, the estimated coefficient on *EDUC* is 0.0609 with standard error 0.005. Adding *ABILITY* to the equation reduces the estimated coefficient on *EDUC* to 0.054 with standard error 0.006. Is this the effect that you anticipate? Explain.

c. Assuming that *ABILITY* and *EXPER* are exogenous, along with instrumental variables *MOTHEREDUC* and *FATHEREDUC*, what is the specification of the first-stage equation? That is, what variables are on the right-hand side?

d. Estimating the first-stage equation in (c), we find that the t-values on *MOTHEREDUC* and *FATHEREDUC* are 2.55 and 4.72, respectively. The F-test of their joint significance is 33.82. Are these instruments adequately strong for their use in IV/2SLS? Explain.

e. Let $\hat{v}$ denote the OLS residuals from part (d). If we estimate the model in Example 10.5, and include the variables *ABILITY* and $\hat{v}$, the t-statistic for $\hat{v}$ is −0.94. What does this result tell us about the endogeneity of *EDUC* after controlling for ability?

10.10 Consider the model in Example 10.5. Suppose we have the idea that the effect of education may differ for individuals who have siblings. Suppose *SIBS* = number of siblings, which we assume is exogenous. We add to the model the variable $EDUC \times SIBS$.

a. Assuming we treat *EDUC* as endogenous, what type of variable is $EDUC \times SIBS$? Is it exogenous or endogenous? Explain your reasoning.

b. In addition to *MOTHEREDUC* and *FATHEREDUC*, are $MOTHEREDUC \times SIBS$ and $FATHEREDUC \times SIBS$ potentially useful IV? Explain how they satisfy, or might satisfy, the three conditions IV1–IV3.

c. Using OLS with a large sample of individuals, we find the estimated coefficient of *EDUC* to be 0.0903 ($t = 46.74$) and the estimated coefficient of $EDUC \times SIBS$ to be −0.0001265 ($t = -0.91$). Explain why we should not simply omit the variable $EDUC \times SIBS$ in the wage equation based on this result.

d. The first-stage equations for *EDUC* and $EDUC \times SIBS$ include *EXPER*, $EXPER^2$, and the four variables listed in (b). The F-tests for the joint significance of the IV have p-values of 0.0000. Can we safely conclude that our IV are strong for both *EDUC* and $EDUC \times SIBS$?

e. We calculate the residuals from the two first-stage equations. Let the residuals from the *EDUC* equation be $\hat{v}_1$ and the residuals from the $EDUC \times SIBS$ equation be $\hat{v}_2$. We estimate the structural model by OLS including both $\hat{v}_1$ and $\hat{v}_2$ as explanatory variables. Their t-values are −10.29 and −1.63, respectively, and the joint F-test of their significance is 55.87. Can we safely conclude that both *EDUC* and $EDUC \times SIBS$ are endogenous?

f. Using IV/2SLS, we find that the estimated coefficient of *EDUC* is 0.1462 with a t-value of 25.25, and the estimated coefficient of $EDUC \times SIBS$ is 0.0007942 with a t-value of 4.53. The estimated covariance between these two coefficients is 4.83×10^{-7}. Estimate the marginal effect of another

year of education on wages for a person with no siblings. What is the estimated marginal effect of education if a person has five siblings?

10.11 Consider the wage equation in Example 10.5.

a. Two possible instruments for *EDUC* are *NEARC4* and *NEARC2*, where these are dummy variables indicating whether the individual lived near a 4-year college or a 2-year college at age 10. Speculate as to why these might be potentially valid IV.

b. Explain the steps (not the computer command) required to carry out the regression-based Hausman test, assuming we use both IV.

c. Using a large data set, the p-value for the regression-based Hausman test for the model in Example 10.5, using only *NEARC4* as an IV is 0.28; using only *NEARC2* the p-value is 0.0736, and using both IV the p-value is 0.0873 [with robust standard errors it is 0.0854]. What should we conclude about the endogeneity of *EDUC* in this model?

d. We compute the IV/2SLS residuals, using both *NEARC4* and *NEARC2* as IV. In the regression of these 2SLS residuals on all exogenous variables and the IV, with $N = 3010$ observations, all regression p-values are greater than 0.30 and the $R^2 = 0.000415$. What can you conclude based on these results?

e. The main reason we seldom use OLS to estimate the coefficients of equations with endogenous variables is that other estimation methods are available that yield better fitting equations. Is this statement true or false, or are you uncertain? Explain the reasoning of your answer.

f. The F-test of the joint significance of *NEARC4* and *NEARC2* in the first-stage regression is 7.89. The 95% interval estimates for the coefficient of education using OLS is 0.0678 to 0.082, and using 2SLS it is 0.054 to 0.260. Explain why the width of the interval estimates is so different.

10.12 Estimating cost and production functions for industrial plants is important. Decisions are based on estimated average and marginal cost, and average and marginal products. Suppose a manufacturing plant for a particular firm has output modeled as $Q = \beta_1 + \beta_2 MGT\_EFF + \beta_3 CAP + \beta_4 LAB + e$, where Q is the output in a particular manufacturing plant, *MGT_EFF* is a managerial efficiency index, *CAP* is capital stock input index and *LAB* is labor input index.

a. What is the interpretation of β_2? What sign should it have?

b. Measuring *MGT_EFF* is difficult. Suppose we propose to estimate the model

$$Q = \beta_1 + \beta_2 XPER + \beta_3 CAP + \beta_4 LAB + e$$

where *XPER* is the plant manager's experience, measured in years. What should the sign of β_2 be now? Why might we worry that *XPER* is endogenous? [*Hint*: Think carefully about this one.]

c. We use data from 75 plants to estimate the model in (b). The least squares estimates are

$$\hat{Q} = 1.7623 + 0.1468\,XPER + 0.4380\,CAP + 0.2392\,LAB$$
$$\text{(se)}\ (1.0550)\ (0.0634)\qquad (0.1176)\qquad (0.0998)$$

Are the signs of the coefficients and their significance consistent with your expectations? Explain.

d. If *XPER* is endogenous, what is the direction of the bias of the OLS estimator? Explain. [*Hint*: Remember your answer to part (b).]

e. Suppose we consider *AGE*, the age of the plant manager, as an instrument. Does it satisfy the criteria for an IV based on your economic reasoning? Why or why not?

f. In the OLS regression of *XPER* on *CAP*, *LAB,* and *AGE,* the t-value for the coefficient of *AGE* is 3.13. What information does this provide us about the feasibility of carrying out IV/2SLS?

g. We add the residuals from part (f) to the model in (b) to obtain

$$Q = \beta_1 + \beta_2 XPER + \beta_3 CAP + \beta_4 LAB + \beta_5 RESID + e$$

The t-statistic for the null hypothesis $H_0: \beta_5 = 0$ from this regression is -2.2. What should we infer from it?

h. The two-stage least squares estimates are

$$\hat{Q} = -2.4867 + 0.5121\,XPER + 0.3321\,CAP + 0.2400\,LAB$$
$$\text{(se)}\ (2.7230)\ (0.2205)\qquad (0.1545)\qquad (0.1209)$$

What are the differences in these estimates versus the OLS estimates? Are the differences consistent with your expectations, relative to the OLS estimates? Explain.

i. Reasoning that *AGE* is an adequate IV, a staff economist decides to add $AGE \times LAB$ and $AGE \times CAP$ as IV also. Are these likely to be valid IV and uncorrelated with the regression error term? To test this, the two-stage least squares residuals are regressed on *CAP*, *LAB*, *AGE*, $AGE \times LAB$, and $AGE \times CAP$. The resulting R^2 is 0.0045. What do you think about the validity of the IV now?

j. The economist regresses *XPER* on *CAP*, *LAB*, *AGE*, $AGE \times LAB$, and $AGE \times CAP$. The *F*-test of the joint significance of *AGE*, $AGE \times LAB$, and $AGE \times CAP$ is 3.3. Do you think it is advisable to use the interaction variables as IV in the estimation? Justify your answer.

10.13 Households plan consumption expenditures and saving with consideration of their long-run income. We wish to estimate $SAVING = \beta_1 + \beta_2 LRINCOME + e$, where *LRINCOME* is long-run income.

a. Long-run income is difficult to define and measure. Using data on 50 households' annual savings (*SAVINGS*, \$1000 units) and annual income (*INCOME*, \$1000 units), we estimate a savings equation by OLS to obtain

$$\widehat{SAVINGS} = 4.3428 - 0.0052 INCOME$$
$$\text{(se)} \qquad (0.8561) \quad (0.0112)$$

Why might we expect the OLS estimator of the marginal propensity to save to be biased and inconsistent? What is the likely direction of the bias?

b. Suppose that in addition to current income we know average household income over the past 10 years (*AVGINC*, \$1000 units). Why might this be a suitable instrumental variable?

c. The estimated first-stage regression is

$$\widehat{INCOME} = -35.0220 + 1.6417 AVGINC$$
$$(t) \qquad (-1.83) \quad (5.80)$$

Does *AVGINC* qualify as a strong instrument? Explain.

d. Let the residuals from part (c) be $\hat{v}$. Adding this variable to the savings equation and estimating the result by OLS gives

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME - 0.0755 \hat{v}$$
$$\text{(se)} \qquad (1.1720) \quad (0.0154) \qquad\qquad (0.0201)$$

Based on this result should we rely on the OLS estimates of the savings equation?

e. Using the fitted values from part (c) in place of *INCOME* and applying OLS, we obtain

$$\widehat{SAVINGS} = 0.9883 + 0.0392 \widehat{INCOME}$$
$$\text{(se)} \qquad (1.2530) \quad (0.0165)$$

Compare these coefficient estimates to those in part (a). Are these estimates more in line with your prior expectations than those in (a), or not?

f. Are the OLS standard errors in part (e) correct or not? Explain.

g. Using IV/2SLS software, with instrument *AVGINC*, we obtain the estimates

$$\widehat{SAVINGS} = 0.9883 + 0.0392 INCOME$$
$$\text{(se)} \qquad (1.5240) \quad (0.0200)$$

Construct a 95% interval estimate of the effect of *INCOME* on *SAVINGS*. Compare and contrast it to the 95% interval estimate based on the results in part (a).

h. In parts (d), (e), and (g), the estimated coefficient of *INCOME* is 0.0392. Is this an accident? Explain.

i. Explain how to test whether *AVGINC* is a valid instrument, and uncorrelated with the regression error.

10.14 The Capital Asset Pricing Model (CAPM) [see Exercise 2.16] says that the risk premium on security *j* is related to the risk premium on the market portfolio, that is

$$r_j - r_f = \alpha_j + \beta_j (r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. We measure the market portfolio using the Standard & Poors value weighted index, and the risk-free rate by the 30-day LIBOR monthly rate of return.

a. Using 180 monthly observations from January 1988, the OLS estimate of IBM's beta is 0.9769 with a standard error of 0.0978. If our constructed values of the market return and the risk-free rate are measured with error is the OLS estimator unbiased and consistent? If it is biased, what is the direction of the bias?

b. It has been suggested that it is possible to construct an IV by ranking the values of the explanatory variable and using the rank as the IV. That is, we sort $(r_m - r_f)$ from smallest to largest, and assign the values $RANK = 1, 2, \ldots, 180$. Does this variable potentially satisfy the conditions IV1–IV3?

c. The estimated first-stage regression of $(r_{IBM} - r_f)$ on *RANK* yields an overall F-test of model significance of 93.77. What can we conclude about the strength of the IV *RANK*?

d. If we compute the first-stage residuals and add them to the CAPM model, the resulting coefficient has a t-value of 60.60. What does this result suggest to us about the OLS estimator in the CAPM model?

e. Using *RANK* as an IV and estimating the CAPM model by IV/2SLS yield an estimate of IBM's beta of 1.0025 with a standard error of 0.1019. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

10.5.2 Computer Exercises

10.15 Consider the simple wage model in Example 10.2. Use the 428 observations on married women who participate in the labor force.

a. Using the instrumental variables estimator in equation (10.17), divide the numerator and denominator by $(N-1)$ and show that the IV estimator is the ratio of sample covariances, $\hat{\beta}_2 = \widehat{\text{cov}}(z_i, y_i)/\widehat{\text{cov}}(z_i, x_i)$.

b. Using your computer software, calculate $\widehat{\text{cov}}(MOTHEREDUC_i, \ln(WAGE_i))$ and $\widehat{\text{cov}}(MOTHEREDUC_i, EDUC_i)$. Compare their ratio to the IV estimate in Example 10.2.

c. In Example 10.5, we added experience and its square to the model specification. To implement the ratio of covariances estimator in part (a), we first remove (partial-out) the influence of experience and its square from *MOTHEREDUC*, *EDUC*, and ln(*WAGE*). Regress each of variables on *EXPER* and $EXPER^2$ and save the residuals, calling them *RMOTHEREDUC*, *REDUC*, and *RLWAGE*. Calculate $\widehat{\text{cov}}(RMOTHEREDUC_i, RLWAGE_i)$ and $\widehat{\text{cov}}(RMOTHEREDUC_i, REDUC_i)$. Compare their ratio to the IV estimate in Example 10.5.

d. Using your IV/2SLS software, estimate the model $RLWAGE = \beta_2 REDUC + error$, omitting the constant term, using *RMOTHEREDUC* as an instrumental variable. Compare the resulting estimate to that in part (c).

10.16 Consider the wage model in Example 10.5 and the 428 observations on married women who participate in the labor force. Use only *MOTHEREDUC* as an instrument in this exercise.

a. Estimate the first-stage equation by OLS and obtain the fitted values

$$\widehat{EDUC} = \hat{\gamma}_1 + \hat{\gamma}_2 EXPER + \hat{\gamma}_3 EXPER^2 + \hat{\theta}_1 MOTHEREDUC$$

b. Use OLS to estimate the second-stage equation

$$\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 \widehat{EDUC} + error$$

c. Obtain the least squares residuals, $\hat{e}$, from the estimation in part (b). Calculate $\sum \hat{e}_i$. Explain why the value you obtain is theoretically correct.

d. Using the coefficient estimates from part (b), calculate the residuals

$$\hat{e}_{IV} = \ln(WAGE) - \hat{\beta}_1 - \hat{\beta}_2 EXPER - \hat{\beta}_3 EXPER^2 - \hat{\beta}_4 EDUC$$

Calculate $\sum \hat{e}_{IV}$. Explain why the value you obtain is theoretically correct.

e. Calculate $\sum \hat{e}_i^2/(N-4)$ and $\sum \hat{e}_{IV}^2/(N-4)$. Which of these is the correct estimator of the error variance, σ^2?

f. Estimate the regression $\widehat{EDUC} = a_1 + a_2 EXPER + a_3 EXPER^2 + error$ and obtain the sum of squared residuals. Use equation (10.25) and one of the values from part (e) to obtain $\widehat{\text{var}}\left(\hat{\beta}_4\right)$.

g. Using software for IV/2SLS estimate the wage model $\ln(WAGE) = \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 EDUC + e$ using the instrumental variable *MOTHEREDUC*. How do the estimates compare to those in part (b)? Does the reported standard error $\text{se}\left(\hat{\beta}_4\right)$ agree with the calculated variance in part (f)?

10.17 Consider the wage model in Example 10.5 and the 428 observations on married women who participate in the labor force. Use only *MOTHEREDUC* as an instrument in this exercise.

a. Estimate the first-stage equation by OLS and obtain the fitted values

$$\widehat{EDUC} = \hat{\gamma}_1 + \hat{\gamma}_2 EXPER + \hat{\gamma}_3 EXPER^2 + \hat{\theta}_1 MOTHEREDUC$$

Save the least squares residuals. Call them *REDUCHAT*. Calculate the sum of squared residuals, $\sum REDUCHAT_i^2$.

b. Estimate the regression $\widehat{EDUC} = a_1 + a_2 EXPER + a_3 EXPER^2 + error$ and save the OLS residuals. Call them *REDUC*. Calculate the sum of squared residuals, $\sum REDUC_i^2$.

c. Estimate the regression $MOTHEREDUC = c_1 + c_2 EXPER + c_3 EXPER^2 + error$ and save the OLS residuals. Call them *RMOM*. Calculate the sum of squared residuals, $\sum RMOM_i^2$.

d. Estimate the regression $REDUC = \theta_1 RMOM + error$. Compare the estimated value of θ_1 from this regression to the estimated θ_1 from the first-stage equation. What R^2 value did you obtain from this regression? What is the sum of squared residuals?

e. Show that $\sum RMOM_i^2 = \hat{\theta}_1^2 \sum REDUC_i^2$.

f. Refer to equation (10.25) and discuss the importance of the quantities in (e) for the precision of the IV/2SLS estimator.

10.18 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of a parent's college education as an instrumental variable.

a. Create two new variables. *MOTHERCOLL* is a dummy variable equaling one if *MOTHEREDUC* > 12, zero otherwise. Similarly, *FATHERCOLL* equals one if *FATHEREDUC* > 12 and zero otherwise. What percentage of parents have some college education in this sample?

b. Find the correlations between *EDUC*, *MOTHERCOLL*, and *FATHERCOLL*. Are the magnitudes of these correlations important? Can you make a logical argument why *MOTHERCOLL* and *FATHERCOLL* might be better instruments than *MOTHEREDUC* and *FATHEREDUC*?

c. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* as the instrumental variable. What is the 95% interval estimate for the coefficient of *EDUC*?

d. For the problem in part (c), estimate the first-stage equation. What is the value of the F-test statistic for the hypothesis that *MOTHERCOLL* has no effect on *EDUC*? Is *MOTHERCOLL* a strong instrument?

e. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* and *FATHERCOLL* as the instrumental variables. What is the 95% interval estimate for the coefficient of *EDUC*? Is it narrower or wider than the one in part (c)?

f. For the problem in part (e), estimate the first-stage equation. Test the joint significance of *MOTHERCOLL* and *FATHERCOLL*. Do these instruments seem adequately strong?

g. For the IV estimation in part (e), test the validity of the surplus instrument. What do you conclude?

10.19 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of a parent's college education as an instrumental variable.

a. Create two new variables. *MOTHERCOLL* is a dummy variable equaling one if *MOTHEREDUC* > 12, zero otherwise. Similarly *FATHERCOLL* equals one if *FATHEREDUC* > 12, and zero otherwise. Also, create *COLLSUM* = *MOTHERCOLL* + *FATHERCOLL* and *COLLBOTH* = *MOTHERCOLL* × *FATHERCOLL*. What values do *COLLSUM* and *COLLBOTH* take? What percentage of women in the sample have both a mother and a father with some college education.

b. Find the correlations between *EDUC*, *COLLSUM*, and *COLLBOTH*. Are the magnitudes of these correlations important? Can you make a logical argument why *COLLSUM* and *COLLBOTH* might be better instruments than *MOTHEREDUC* and *FATHEREDUC*?

c. Estimate the wage equation in Example 10.5 using 2SLS with *COLLSUM* as the instrumental variable. What is the 95% interval estimate for the coefficient of *EDUC*?

d. For the problem in part (c), estimate the first-stage equation. What is the value of the F-test statistic for the hypothesis that *COLLSUM* has no effect on *EDUC*? Is *COLLSUM* a strong instrument?

e. Using OLS estimate the regression model with *EDUC* as dependent variable, and include as explanatory variables experience, and its square, along with *MOTHERCOLL* and *FATHERCOLL*, and a constant term. Test the null hypothesis that the coefficients of *MOTHERCOLL* and *FATHERCOLL* are equal at the 5% level.

f. Based on the results in part (e), are we justified in using *COLLSUM* = *MOTHERCOLL* + *FATHERCOLL* as an IV? Are we better off using *COLLSUM* only or using *MOTHERCOLL* and *FATHERCOLL*?

10.20 The CAPM [see Exercises 10.14 and 2.16] says that the risk premium on security j is related to the risk premium on the market portfolio. That is

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. We measure the market portfolio using the Standard & Poor's value weighted index, and the risk-free rate by the 30-day LIBOR monthly rate of return. As noted in Exercise 10.14, if the market return is measured with error, then we face an errors-in-variables, or measurement error, problem.

a. Use the observations on Microsoft in the data file *capm5* to estimate the CAPM model using OLS. How would you classify the Microsoft stock over this period? Risky or relatively safe, relative to the market portfolio?

b. It has been suggested that it is possible to construct an IV by ranking the values of the explanatory variable and using the rank as the IV, that is, we sort $(r_m - r_f)$ from smallest to largest, and assign the values $RANK = 1, 2, \ldots, 180$. Does this variable potentially satisfy the conditions IV1–IV3? Create *RANK* and obtain the first-stage regression results. Is the coefficient of *RANK* very significant? What is the R^2 of the first-stage regression? Can *RANK* be regarded as a strong IV?

c. Compute the first-stage residuals, $\hat{v}$, and add them to the CAPM model. Estimate the resulting augmented equation by OLS and test the significance of $\hat{v}$ at the 1% level of significance. Can we conclude that the market return is exogenous?

d. Use *RANK* as an IV and estimate the CAPM model by IV/2SLS. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

e. Create a new variable $POS = 1$ if the market return $(r_m - r_f)$ is positive, and zero otherwise. Obtain the first-stage regression results using both *RANK* and *POS* as instrumental variables. Test the joint significance of the IV. Can we conclude that we have adequately strong IV? What is the R^2 of the first-stage regression?

f. Carry out the Hausman test for endogeneity using the residuals from the first-stage equation in (e). Can we conclude that the market return is exogenous at the 1% level of significance?

g. Obtain the IV/2SLS estimates of the CAPM model using *RANK* and *POS* as instrumental variables. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

h. Obtain the IV/2SLS residuals from part (g) and use them (not an automatic command) to carry out a Sargan test for the validity of the surplus IV at the 5% level of significance.

10.21 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of alternative constructed instrumental variables. Estimate the model in Example 10.5 using IV/2SLS using both *MOTHEREDUC* and *FATHEREDUC* as IV. These will serve as our baseline results.

a. Write down the first-stage equation using econometric notation, as in equation (10.26), with $\gamma_1, \gamma_2, \gamma_3$ as the unknown coefficients of the intercept, *EXPER* and its square, and θ_1, θ_2 as the coefficients of *MOTHEREDUC* and *FATHEREDUC*, respectively. Test the null hypothesis that $\theta_1 = \theta_2$ at the 5% level. What do you conclude?

b. Assume that $\theta_1 = \theta_2 = \theta$. Substitute into the first-stage equation to obtain a "restricted" model. What variable involving *MOTHEREDUC* and *FATHEREDUC* now appears on the right-hand side?

c. Create a new variable $PARENTSUM = MOTHEREDUC + FATHEREDUC$. Obtain IV/2SLS estimates using this as the IV. How do the estimates compare to the baseline results? Is this IV strong?

d. Create two new variables $MOMED2 = MOTHEREDUC^2$ and $DADED2 = FATHEREDUC^2$. Use these new variables and both *MOTHEREDUC* and *FATHEREDUC* as IV. Estimate the first-stage equation using these four IV. Test their joint significance using an F-test. Are these instruments adequately strong? Do any seem irrelevant based on t-tests of significance? Find the simple correlations among the four IV. Are any large?

e. Obtain IV/2SLS estimates of the model in Example 10.5 using the four IV in part (d). How do these estimates compare to the baseline results and to those in part (c)?

f. Based on the results in this question, which set of IV/2SLS estimates would you prefer to report? The baseline estimates, the results in part (c), or the results in part (e). Explain your choice.

10.22 Consider the data file *mroz* on working wives and the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$. Use the 428 observations on married women who participate in the labor force.

a. Write down in algebraic form the three moment conditions, like (10.13) and (10.14), that would lead to the OLS estimates of the model above.

b. Calculate the OLS estimates and residuals, $\hat{e}_i$. What is the sum of the least squares residuals? What is the sum of squared least squares residuals? What is $\sum EDUC_i \times \hat{e}_i$? What is $\sum EXPER_i \times \hat{e}_i$? Relate these results to the moment conditions in (a).

c. Calculate the fitted values $\widehat{\ln(WAGE)} = b_1 + b_2 EDUC + b_3 EXPER$. What is the sample average of the fitted values? What is the sample average of $\ln(WAGE)$, $\overline{\ln(WAGE)}$?

d. Find each of the following:

$$SST = \sum\left[\ln(WAGE_i) - \overline{\ln(WAGE)}\right]^2, \quad SSE = \sum \hat{e}_i^2, \quad SSR = \sum\left[\widehat{\ln(WAGE_i)} - \overline{\ln(WAGE)}\right]^2$$

Compute $SSR + SSE$, $R^2 = SSR/SST$ and $R^2 = 1 - SSE/SST$. Explain what these calculations show about measuring goodness-of-fit.

10.23 This question is an extension of Exercise 10.22. Consider the data file *mroz* on working wives and the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$. Use the 428 observations on married women who participate in the labor force. Let the instrumental variable be *MOTHEREDUC*.

a. Write down in algebraic form the three moment conditions, like (10.16) , that would lead to the IV/2SLS estimates of the model above.

b. Calculate the IV/2SLS estimates and residuals, $\hat{e}_{IV}$. What is the sum of the IV residuals? What is $\sum MOTHEREDUC_i \times \hat{e}_{IV,i}$? What is $\sum EXPER_i \times \hat{e}_{IV,i}$? Relate these results to the moment conditions in (a).

c. What is $\sum EDUC_i \times \hat{e}_{IV,i}$? What is the sum of squared IV residuals? How do these two results compare with the corresponding OLS results in Exercise 10.22(b)?

d. Calculate the IV/2SLS fitted values $FLWAGE = \hat{\beta}_1 + \hat{\beta}_2 EDUC + \hat{\beta}_3 EXPER$. What is the sample average of the fitted values? What is the sample average of $\ln(WAGE)$, $\overline{\ln(WAGE)}$?

e. Find each of the following:

$$SST = \sum\left[\ln(WAGE_i) - \overline{\ln(WAGE)}\right]^2, \quad SSE\_IV = \sum \hat{e}_{IV}^2,$$
$$SSR\_IV = \sum\left[FLWAGE - \overline{\ln(WAGE)}\right]^2$$

Compute $SSR\_IV + SSE\_IV$, $R^2_{IV,1} = SSR\_IV/SST$, and $R^2_{IV,2} = 1 - SSE\_IV/SST$. How do these values compare to those in Exercise 10.22(d)?

f. Does your IV/2SLS software report an R^2 value. Is it either of the ones in (e)? Explain why the usual concept of R^2 fails to hold for IV/2SLS estimation.

10.24 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of alternative standard errors for the IV estimator. Estimate the model in Example 10.5 using IV/2SLS using both *MOTHEREDUC* and *FATHEREDUC* as IV. These will serve as our baseline results.

a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity?

b. Regress $\hat{e}_{IV}^2$ against a constant and *EXPER*. Apply the NR^2 test from Chapter 8 to test for the presence of heteroskedasticity.

c. Obtain the IV/2SLS estimates with the software option for Heteroskedasticity Robust Standard Errors. Are the robust standard errors larger or smaller than those for the baseline model? Compute the 95% interval estimate for the coefficient of *EDUC* using the robust standard error.

d. Obtain the IV/2SLS estimates with the software option for Bootstrap standard errors, using $B = 200$ bootstrap replications. Are the bootstrap standard errors larger or smaller than those for the baseline model? How do they compare to the heteroskedasticity robust standard errors in (c)? Compute the 95% interval estimate for the coefficient of *EDUC* using the bootstrap standard error.

10.25 To examine the quantity theory of money, Brumm (2005) ["Money Growth, Output Growth, and Inflation: A Reexamination of the Modern Quantity Theory's Linchpin Prediction," *Southern Economic Journal*, 71(3), 661–667] specifies the equation

$$INFLATION = \beta_1 + \beta_2 MONEY\ \ GROWTH + \beta_3 OUTPUT\ \ GROWTH + e$$

where *INFLATION* is the growth rate of the general price level, *MONEY GROWTH* is the growth rate of the money supply, and *OUTPUT GROWTH* is the growth rate of national output. According to theory we should observe that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$. Use the data file *brumm*. It consists of 1995 data on 76 countries. We wish to test

i. the *strong* joint hypothesis that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$.

ii. the *weak* joint hypothesis $\beta_2 = 1$ and $\beta_3 = -1$

a. It is argued that *OUTPUT GROWTH* may be endogenous. Four instrumental variables are proposed, *INITIAL* = initial level of real GDP, *SCHOOL* = a measure of the population's educational attainment, *INV* = average investment share of GDP, and *POPRATE* = average population growth rate. Using these instruments, obtain instrumental variables (2SLS) estimates of the inflation equation.

b. Test the strong and weak hypotheses using the IV estimates.

c. Compute the IV/2SLS residuals, $\hat{e}_{IV}$. Identify the observation with the largest absolute residual, $|\hat{e}_{IV}|$. How does it compare to the next smallest residual?

d. Let us examine the effect of the observation with the largest residual. Drop the corresponding observation from the data, reestimate the model using IV/2SLS, and carry out the tests of the strong and weak hypotheses. How much do things change, if any?

e. Obtain the IV/2SLS residuals from part (d), $\tilde{e}_{IV}$. Regress $\tilde{e}_{IV}^2$ on *MONEY*. Calculate the heteroskedasticity test statistic NR^2. Compare it to the 95th percentile of the $\chi^2_{(1)}$ distribution. Is there evidence of heteroskedasticity?

f. Using the 75 remaining observations from (d) obtain the IV/2SLS estimates with heteroskedasticity robust standard errors. Carry out the tests of the strong and weak hypotheses. How to the test results compare to those in (d)?

g. Using the remaining 75 observations from (d), estimate the first-stage equation and test the joint significance of the IV. Repeat the tests robust to heteroskedasticity. Is there evidence that the instruments are strong?

h. Regress $\tilde{e}_{IV}$ against the four IV and *MONEY*. Are any of the coefficients significant? If the IV are valid, do we expect any significant coefficients in this regression? Explain.

Appendix 10A Testing for Weak Instruments

The F-test for weak instruments discussed in Section 10.3.9 is not valid for models with more than one endogenous variable on the right side of the equation.[7] Using **canonical correlations** there is a solution to the problem of identifying weak instruments when an equation has more than one endogenous variable. Canonical correlations are a generalization of the usual concept of

[7] The $F > 10$ rule of thumb comes from D. Staiger and J.H. Stock (1997) "Instrumental Variables with Weak Instruments," *Econometrica* 65, pp. 557–586.

a correlation between two variables and attempt to describe the association between two **sets** of variables. The association in which we are interested is the association between the pair of endogenous variables (x_{G+1}, x_{G+2}) and the pair of additional, external, instrumental variables (z_1, z_2) **after** controlling for the effect of the other G exogenous variables $\mathbf{x}_1 \equiv (1, x_2, \ldots, x_G)$. The effects of the G exogenous variables are "removed" by first regressing (x_{G+1}, x_{G+2}) and (z_1, z_2) on $\mathbf{x}_1$ and then computing the residuals $(\tilde{x}_{G+1}, \tilde{x}_{G+2})$ and $(\tilde{z}_1, \tilde{z}_2)$. This process is often called **partialing out** the effect of $\mathbf{x}_1$.

Suppose that $x_1^* = h_{11}\tilde{x}_{G+1} + h_{21}\tilde{x}_{G+2}$ is a linear combination of the "partialed out" endogenous variables $(\tilde{x}_{G+1}, \tilde{x}_{G+2})$ and $z_1^* = k_{11}\tilde{z}_1 + k_{21}\tilde{z}_2$ is a linear combination of the "partialed out" instrumental variables $(\tilde{z}_1, \tilde{z}_2)$. Using **canonical correlation analysis**, we can determine values h_{11}, h_{21}, k_{11}, and k_{21}, resulting in the largest correlation between x_1^* and z_1^*.[8] It is called the **first** canonical correlation, r_1. Similarly, we can determine values h_{12}, h_{22}, k_{12}, and k_{22}, resulting in the second largest correlation between $x_2^* = h_{12}\tilde{x}_{G+1} + h_{22}\tilde{x}_{G+2}$ and $z_2^* = k_{12}\tilde{z}_1 + k_{22}\tilde{z}_2$, which is called the **second** canonical correlation, r_2—and so on.

If we have two variables in the first set of variables and two variables in the second set, then there are two canonical correlations, r_1 and r_2. If we have B variables in the first group (the endogenous variables with the effects of $\mathbf{x}_1$ removed) and $L \geq B$ variables in the second group (the group of instruments with the effects of $\mathbf{x}_1$ removed), then there are B possible canonical correlations, $r_1 \geq r_2 \geq \cdots \geq r_B$. If the **smallest** canonical correlation $r_B = 0$, then we do not have enough relationships between the instruments and the endogenous variables, and **the equation is not identified**.

10A.1 A Test for Weak Identification

Using the smallest canonical correlation, we are able to test whether any relationship between the instruments and the endogenous variables is sufficiently strong for reliable econometric inferences.[9] Let N denote the sample size, B the number of right-hand side endogenous variables, G the number of exogenous variables included in the equation (including the intercept), L the number of "external" instruments that are not included in the model, and r_B the minimum canonical correlation. A test for weak identification, the situation that arises when the instruments are correlated with the endogenous regressors but only weakly, is based on the **Cragg–Donald *F*-test statistic**[10]

$$\text{Cragg–Donald } F = \left[(N-L)/L\right] \times \left[r_B^2/\left(1-r_B^2\right)\right] \tag{10A.1}$$

The Cragg–Donald statistic reduces to the usual weak instruments F-test when the number of endogenous variables is $B = 1$. Critical values for this test statistic have been tabulated by James Stock and Motohiro Yogo (2005),[11] so that we can test the null hypothesis that the instruments

[8] Certain normalizations on h and k constants are necessary to make the solutions unique. The algebra and calculations are beyond the scope of this book. An online search will reveal many sources but virtually all use matrix algebra and multidimensional calculus. Harold Hotelling did research in mathematical statistics and economic theory and introduced the concept of canonical correlation in a 1935 publication."The most predictable criterion," in the *Journal of Educational Psychology*.

[9] The tests based on canonical correlations are neatly summarized in "Enhanced Routines for Instrumental Variables/ Generalized Method of Moments Estimation and Testing," by Christopher F. Baum, Mark E. Schaffer, and Steven Stillman, *The Stata Journal* (2007), 7, pp. 465–506. Further discussion is provided by Alastair R. Hall, Glenn D. Rudebusch and David W. Wilcox (1996) "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37(2), pp. 283–298.

[10] Cragg, J. G. and S. G. Donald (1993) "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory,* 9, 222–240. D. Poskitt and C. Skeels (2009), "Assessing the Magnitude of the Concentration Parameter in a Simultaneous Equations Model." *The Econometrics Journal,* 12, pp. 26–44, showed that the Cragg–Donald statistic could be conveniently written in terms of the smallest canonical correlation.

[11] "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds, Donald W. K. Andrews and James H. Stock. Cambridge University Press, Chapter 5.

are weak against the alternative that they are not, for two particular consequences of weak instruments.

- **Relative Bias:** In the presence of weak instruments, the amount of bias in the IV estimator can become large. Stock and Yogo consider the bias when estimating the coefficients of the endogenous variables. They examine the maximum IV estimator bias relative to the bias of the least squares estimator. Stock and Yogo give the illustration of estimating the return to education. If a researcher believes that the least squares estimator suffers a maximum bias of 10%, and if the relative bias is 0.1, then the maximum bias of the IV estimator is 1%.
- **Rejection Rate (Test Size):** When estimating a model with endogenous regressors, testing hypotheses about the coefficients of the endogenous variables is frequently of interest. If we choose the $\alpha = 0.05$ level of significance, we expect that a true null hypothesis is rejected 5% of the time in repeated samples. If instruments are weak, then the actual rejection rate of the null hypothesis, also known as the **test size**, may be larger. Stock and Yogo's second criterion is the maximum rejection rate of a true null hypothesis if we choose $\alpha = 0.05$. For example, we may be willing to accept a maximum rejection rate of 10% for a test at the 5% level, but we may not be willing to accept a rejection rate of 20% for a 5% level test.

To test the null hypothesis that instruments are weak against the alternative that they are not, we compare the Cragg–Donald F-test statistic to a critical value chosen from Table 10A.1 or Table 10A.2.

TABLE 10A.1 **Critical Values for the Weak Instrument Test Based on IV Test Size (5% level of significance)[12]**

	$B = 1$ Maximum Test Size				$B = 2$ Maximum Test Size			
L	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	19.93	11.59	8.75	7.25	7.03	4.58	3.95	3.63
3	22.30	12.83	9.54	7.80	13.43	8.18	6.40	5.45
4	24.58	13.96	10.26	8.31	16.87	9.93	7.54	6.28

TABLE 10A.2 **Critical Values for the Weak Instrument Test Based on IV Relative Bias (5% level of significance)[13]**

	$B = 1$ Maximum Relative Bias				$B = 2$ Maximum Relative Bias			
L	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
3	13.91	9.08	6.46	5.39				
4	16.85	10.27	6.71	5.34	11.04	7.56	5.57	4.73

[12]These values are from Table 5.2, page 101, in Stock and Yogo (2005), *op cit*. The authors thank James Stock and Motohiro Yogo for permission to use these results. (Their tables are more extensive than the ones we provide.)

[13]These values are from Table 5.1, page 100, in James H. Stock and Motohiro Yogo (2005), *op cit*. In their paper Stock and Yogo explain that the $F > 10$ rule introduced by Staiger and Stock (1997), op cit., is for $B = 1$ approximately the critical value for a maximum relative bias of 0.10 for all values of L. Their critical values can be considered refinements of the Staiger–Stock rule of thumb.

1. **First** choose either the maximum relative bias or maximum test size criterion. You must also choose the maximum relative bias or maximum test size you are willing to accept.

2a. If you choose the maximum test size criterion, select from Table 10A.1 the critical value associated with a maximum test size of 0.10, 0.15, 0.20, or 0.25 for $B = 1$ or $B = 2$ endogenous variables using $L = 1$ to $L = 4$ instrumental variables.

2b. If you choose the maximum relative bias criterion, select from Table 10A.2 the critical value associated with a maximum relative bias of 0.05, 0.10, 0.20, or 0.30 for $B = 1$ or $B = 2$ endogenous variables using $L = 3$ or $L = 4$ instrumental variables. There are no critical values using this criterion if $L < 3$.

3. Reject the null hypothesis that the instruments are weak if the Cragg–Donald F-test statistic is larger than the tabled critical value. If the F-test statistic is not larger than the critical value, then do not reject the null hypothesis that the instruments are weak.

EXAMPLE 10.8 | Testing for Weak Instruments

In Section 10.2.4 we introduced an example of a wage equation for married working women using Thomas Mroz's data. Consider the following *HOURS* supply equation specification:

$$HOURS = \beta_1 + \beta_2 MTR + \beta_3 EDUC + \beta_4 KIDSL6 + \beta_5 NWIFEINC + e \quad (10A.4)$$

The variable $NWIFEINC = (FAMINC - WAGE \times HOURS)/1000$ is household income attributable to sources other than the wife's income. The variable *MTR* is the marginal tax rate facing the wife, including Social Security taxes. In this equation we expect the signs of coefficients on *MTR*, *KIDSL6*, and *NWIFEINC* to be negative, and the coefficient on *EDUC* is of uncertain sign. In this example, we treat the marginal tax rate as endogenous.[14] Initially we treat *EDUC* as exogenous and use the wife's previous years of work experience, *EXPER*, as an instrumental variable for *MTR*.

Weak IV Example 1: Endogenous: *MTR*; Instrument: *EXPER*

Suppose that we choose the maximum test size criterion and are willing to accept a maximum test size of 0.15 for a 5% test. In Table 10A.1, we see that for $B = 1$ (one right-hand side endogenous variable) and $L = 1$ (one instrument) that the Stock-Yogo critical value is 8.96. The estimated first-stage equation for *MTR* is Model (1) of Table 10A.3. The F-statistic for the hypothesis that the coefficient of experience is zero is 30.61. The Cragg–Donald F-statistic is also 30.61 in this case. Since the Cragg–Donald F-test statistic is larger than the Stock-Yogo critical value 8.96, we reject the null hypothesis that the instruments are weak and accept the alternative that they are not weak. This conclusion is conditional upon the test criterion we have chosen and the maximum size selected. The relative bias criterion cannot be used in this case because it requires at least three instruments. The estimated coefficient of *MTR* in the estimated *HOURS* supply equation in Model (1) of Table 10A.4 is negative and significant at the 5% level.

Weak IV Example 2: *Endogenous: MTR; Instruments: EXPER, EXPER*^{2}*, LARGECITY*

For the sake of illustration, consider using the $L = 3$ instruments *EXPER*, *EXPER*^{2}, and the indicator variable *LARGECITY*, which = 1 if the city is large. Suppose we choose the maximum relative bias criterion and are willing to tolerate a maximum relative bias of 0.10. From Table 10A.2 the Stock–Yogo critical value is 9.08. If the Cragg–Donald F-test statistic is greater than this value, we reject the null hypothesis that the instruments are weak. The first-stage equation estimates are reported in Model (2) of Table 10A.3. The Cragg–Donald F-statistic is 13.22. We conclude that using this test the instruments are not weak. If, however, we are only willing to accept a 0.05 relative bias, then the Stock–Yogo critical value is 13.91. Since the Cragg–Donald F-statistic is less than this value, we cannot reject the null hypothesis that the instruments are weak. The estimated coefficient of *MTR* in the estimated *HOURS* supply equation in Model (2) of Table 10A.4 is negative and significant at the 5% level, although the magnitudes of all the coefficients are smaller in absolute value for this estimation than for the model in Model (1). Qualitatively the estimates of Model (1) and Model (2), using $L = 1$ instrument and $L = 3$ instruments are much the same, with likely thanks to the strong instrument *EXPER*. This example illustrates the point that having more instrumental variables is not necessarily beneficial from the standpoint of weak instrument diagnostics.

[14] This idea is explored by Mroz (1987, p. 786).

TABLE 10A.3 First-stage Equations

MODEL Dependent/ independent	(1) *MTR*	(2) *MTR*	(3) *MTR*	(4) *EDUC*	(5) *MTR*	(6) *EDUC*
C	0.87930 (74.33)	0.88470 (71.93)	0.79907 (103.22)	8.71459 (25.83)	0.82960 (93.34)	8.17622 (20.34)
EXPER	−0.00142 (−5.53)	−0.00217 (−2.65)			−0.00168 (−6.23)	0.02957 (2.43)
EDUC	−0.00718 (−7.76)	−0.00689 (−7.45)				
KIDSL6	0.02037 (3.86)	0.02039 (3.89)	0.02189 (3.92)	0.61812 (2.54)	0.01559 (2.87)	0.72921 (2.96)
NWIFEINC	−0.00551 (−27.40)	−0.00539 (−26.35)	−0.00565 (−27.15)	0.04961 (5.46)	−0.00585 (−28.96)	0.05304 (5.81)
$EXPER^2$		0.00002 (1.01)				
LARGECITY		−0.01163 (−2.70)				
MOTHEREDUC			−0.00111 (−1.40)	0.15202 (4.40)	−0.00134 (−1.76)	0.15601 (4.54)
FATHEREDUC			−0.00180 (−2.40)	0.16371 (5.01)	−0.00202 (−2.81)	0.16754 (5.15)
N	428	428	428	428	428	428
Weak IV *F*	30.61	13.22	8.14	49.02	18.86	35.03
Number IV *L*	1	3	2	2	3	3
Number Endog *B*	1	1	2	2	2	2

t-statistics in parentheses.

TABLE 10A.4 IV Estimation of Hours Equation

MODEL	(1)	(2)	(3)	(4)
C	17423.7211 (5.56)	14394.1144 (5.68)	−24491.5995 (−0.31)	18067.8425 (5.11)
MTR	−18456.5896 (−5.08)	−14934.3696 (−5.09)	29709.4677 (0.33)	−18633.9223 (−4.85)
EDUC	−145.2928 (−4.40)	−118.8846 (−4.28)	258.5590 (0.32)	−189.8611 (−3.04)
KIDSL6	151.0229 (1.07)	58.7879 (0.48)	−1144.4779 (−0.46)	190.2755 (1.20)
NWIFEINC	−103.8983 (−5.27)	−85.1934 (−5.32)	149.2325 (0.31)	−102.1516 (−5.11)
N	428	428	428	428
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60

t-statistics in parentheses.

Weak IV Example 3 *Endogenous: MTR, EDUC; Instruments: MOTHEREDUC, FATHEREDUC*

Now treat both marginal tax rate *MTR* and education *EDUC* as endogenous, so that $B = 2$. Following Section 10.3.6 we use mother's and father's education, *MOTHEREDUC* and *FATHEREDUC*, as instruments, so that $L = 2$. Suppose that we are willing to accept a maximum test size of 15% for a 5% test. From Table 10A.1 the critical value for the weak instrument test is 4.58. The first-stage equations for *MTR* and *EDUC* are Model (3) and Model (4) of Table 10A.3. These instruments are strong for *EDUC* as we have earlier seen, with the first-stage weak instrument F-test statistic 49.02. For *MTR* [Model (3)] these two instruments are less strong. *FATHEREDUC* is significant at the 5% level, and the first-stage weak instrument F-test statistic is 8.14, which has a p-value of 0.0003. While this does not satisfy the $F \geq 10$ rule of thumb, it is "close," and we may have concluded that these two instruments were adequately strong. The Cragg–Donald F-test statistic value is only 0.101, which is far below the critical value 4.58 for 15% maximum test size (for a 5% test on *MTR* and *EDUC*). We cannot reject the null hypothesis that the instruments are *weak*, despite the favorable first-stage F-test values. The estimates of the *HOURS* supply equation, Model (3) of Table 10A.3, shows parameter estimates that are wildly different from those in Model (1) and Model (2), and the very small t-statistic values imply very large standard errors, another consequence of instrumental variables estimation in the presence of weak instruments.

Weak IV Example 4 *Endogenous: MTR, EDUC; Instruments: MOTHEREDUC, FATHEREDUC, EXPER*

If we include the additional instrument *EXPER*, so that $L = 3$, we obtain the first-stage estimates in Model (5) and Model (6) of Table 10A.3. Once again the first-stage weak instrument F-test statistic values appear strong, with values for *MTR* of 18.86 and for *EDUC* of 35.03. Using the $F > 10$ rule of thumb, we would be comfortable that our instruments are strong. The Cragg–Donald F-test statistic value is 8.60, which tells a slightly different story. Our instruments are not quite as strong as the first-stage weak instrument F-test statistics imply. If we choose a maximum test size of 0.15, we can reject the null hypothesis of weak instruments. If, however, we are prepared to accept only a maximum 10% rejection rate for a 5% test, the critical value is 13.43, and we do not reject the null hypothesis that the instruments are weak. The instrumental variables estimates of the *HOURS* supply equation are Model (4) of Table 10A.4 and we see that they are more in line with Model (1) and Model (2) than those in Model (3).

10A.2 Testing for Weak Identification: Conclusions

If instrumental variables are "weak," then the instrumental variables, or two-stage least squares, estimator is unreliable. When there is a single endogenous variable, the first-stage F-test of the joint significance of the external instruments is an indicator of instrument strength. The $F > 10$ rule of thumb has been refined by Stock and Yogo, who provide tables of critical values for the null hypothesis "the instruments are weak" using two criteria: the bias of the IV estimator relative to the bias of the least squares estimator, and the maximum size of a 5% test of the coefficients of the endogenous variables. If there is more than one endogenous variable on the right-hand side of an equation, then the F-test statistics from the first-stage equations do not provide reliable information about instrument strength. In this case the Cragg–Donald F-test statistic should be used to test for weak instruments, along with the Stock-Yogo tables of critical values.

Econometric research continues for alternatives to the IV/2SLS estimator in the weak instrument case. Some progress has been made; these results are summarized in Appendix 11B. The discussion is deferred until the next chapter, as the advances have their genesis in discussions of estimation of simultaneous equations models.

Appendix 10B Monte Carlo Simulation

In this appendix we do two sorts of simulations. First, we generate a sample of artificial data and give numerical illustrations of the estimators and tests discussed in the chapter. In the chapter the illustrations used real data. The advantage gained here is that we can see how the estimators and tests perform using data we know comes from a particular data generation process. Secondly, we carry out a Monte Carlo simulation to illustrate the repeated **sampling properties** of the least squares and IV/2SLS estimators under various conditions.

10B.1 Illustrations Using Simulated Data

In this section, we demonstrate, using a simulated sample of data, that the OLS estimator fails when $\text{cov}(x_i, e_i) \neq 0$, and that instrumental variables estimators "work" when conditions listed in Section 10.3.3 are satisfied. For the simulated data, we specify a simple regression model in which the parameter values are $\beta_1 = 1$ and $\beta_2 = 1$. Thus, the systematic part of the regression model is $E(y|x) = \beta_1 + \beta_2 x = 1 + 1 \times x$. By adding to $E(y|x)$ an error term value, which will be a random number we create, we can create a sample value of y.

We want to explore the properties of the OLS estimator when x and e are correlated. Using random number generators, we create $N = 100$ pairs of x and e values, such that each has a normal distribution with mean zero and variance one. The population correlation between the x and e values is ρ_{xe}. We then create an artificial sample of y values by adding e to the systematic portion of the regression,

$$y = E(y|x) + e = \beta_1 + \beta_2 x + e = 1 + 1 \times x + e$$

The data values are contained in the data file *ch10*. The OLS estimates are

$$\begin{aligned} \hat{y}_{\text{OLS}} &= 0.9789 + 1.7034x \\ (\text{se}) \quad & (0.088) \quad (0.090) \end{aligned}$$

When x and e are positively correlated, the estimated slope tends to be too large—here, $b_2 = 1.7034$ compared to the true $\beta_2 = 1$. Furthermore, the systematic overestimation of the slope will not go away in larger samples, so the least squares estimators are not correct on average even in large samples. The least squares estimators are inconsistent.

In the process of creating the artificial data (data file *ch10*) we also created two instrumental variables, both uncorrelated with the error term. The correlation between the first instrument z_1 and x is $\rho_{xz_1} = 0.5$, and the correlation between the second instrument z_2 and x is $\rho_{xz_2} = 0.3$. The IV estimates using z_1 are

$$\begin{aligned} \hat{y}_{IV\_z_1} &= 1.1011 + 1.1924x \\ (\text{se}) \quad & (0.109) \quad (0.195) \end{aligned}$$

and the IV estimates using z_2 are

$$\begin{aligned} \hat{y}_{IV\_z_2} &= 1.3451 + 0.1724x \\ (\text{se}) \quad & (0.256) \quad (0.797) \end{aligned}$$

Using z_1, the stronger instrument, yields an estimate of the slope of 1.1924 with a standard error of 0.195, about twice the standard error of the OLS estimate. Using the weaker instrument z_2 produces a slope estimate of 0.1724, which is far from the true value, and a standard error of 0.797, about eight times as large as the least squares standard error. The results with the weaker instrument are far less satisfactory than the estimates based on the stronger instrument z_1.

Another problem that an instrument can have is that it is not uncorrelated with the error term as it is supposed to be. The variable z_3 is correlated with x, with correlation $\rho_{xz_3} = 0.5$, but it is correlated with the error term e, with correlation $\rho_{ez_3} = 0.3$. Thus, z_3 is not a valid instrument. What happens if we use instrumental variables estimation with the invalid instrument? The results are

$$\begin{aligned} \hat{y}_{IV-z_3} &= 0.9640 + 1.7657x \\ (\text{se}) \quad & (0.095) \quad (0.172) \end{aligned}$$

As you can see, using the invalid instrument produces a slope estimate even further from the true value than the least squares estimate. Using an invalid instrumental variable means that the instrumental variables estimator will be inconsistent, just like the least squares estimator.

What is the outcome of two-stage least squares estimation using the two instruments z_1 and z_2? Obtain the first-stage regression of x on the two instruments z_1 and z_2,

$$\begin{aligned} \hat{x} &= 0.1947 + 0.5700z_1 + 0.2068z_2 \\ (\text{se}) \ (0.095) \quad &(0.089) \quad (0.077) \end{aligned} \tag{10B.1}$$

Using the predicted value $\hat{x}$ to replace x, then applying least squares to the modified equation, as in (10.22), we obtain the instrumental variables estimates

$$\begin{aligned} \hat{y}_{IV\_z_1,z_2} &= 1.1376 + 1.0399x \\ (\text{se}) \quad &(0.116) \quad (0.194) \end{aligned} \tag{10B.2}$$

The standard errors are based on an estimated error variance as in (10.18b). Using the two valid instruments yields an estimate of the slope of 1.0399, which, in this example, is close to the true value of $\beta_2 = 1$.

10B.1.1 The Hausman Test

To implement the Hausman test we estimate the first-stage equation, which is shown in (10A.1) using the instruments z_1 and z_2. Compute the residuals

$$\hat{v} = x - \hat{x} = x - 0.1947 - 0.5700z_1 - 0.2068z_2$$

Include the residuals as an extra variable in the regression equation and apply least squares,

$$\begin{aligned} \hat{y} &= 1.1376 + 1.0399x + 0.9957\hat{v} \\ (\text{se}) \ (0.080) \quad &(0.133) \quad (0.163) \end{aligned}$$

The t-statistic for the null hypothesis that the coefficient of $\hat{v}$ is zero is 6.11. The critical value comes from the t-distribution with 97 degrees of freedom and is 1.985, so we reject the null hypothesis that x is uncorrelated with the error term and correctly conclude that it is endogenous.

10B.1.2 Test for Weak Instruments

The test for weak instruments again begins with estimation of the first-stage regression. If we consider using just z_1 as an instrument, the estimated first-stage equation is

$$\begin{aligned} \hat{x} &= 0.2196 + 0.5711z_1 \\ (t) \quad &(6.24) \end{aligned}$$

The t-statistic 6.24 corresponds to an F-value of 38.92, which is well above the guideline value of 10. If we use just z_2 as an instrument, the estimated first-stage equation is

$$\begin{aligned} \hat{x} &= 0.2140 + 0.2090z_2 \\ (t) \quad &(2.28) \end{aligned}$$

While the t-statistic 2.28 indicates statistical significance at the 0.05 level, the corresponding F-value is $5.21 < 10$, indicating that z_2 is a weak instrument. The first-stage equation using both instruments is shown in (10B.1), and the F-test for their joint significance is 24.28, indicating that we have at least one strong instrument.

10B.1.3 Testing the Validity of Surplus Instruments

If we use z_1 and z_2 as instruments, there is one extra. The number of instruments is $L = 2$, and the number of endogenous regressors is $B = 1$. The IV estimates are shown in (10B.2). Calculate the residuals from this equation and then regress them on intercept, z_1 and z_2, to obtain $\hat{e} = 0.0189 + 0.0881z_1 - 0.1818z_2$. The R^2 from this regression is 0.03628, and $NR^2 = 3.628$. The 0.05 critical value for the chi-square distribution with one degree of freedom is 3.84, so we fail to reject the validity of the surplus moment condition.

If we use z_1, z_2, and z_3 as instruments, there are two surplus moment conditions. The IV estimates using these three instruments are $\hat{y}_{IV\_z_1,z_2,z_3} = 1.0626 + 1.3535x$. Obtaining the residuals and regressing them on the instruments yields

$$\hat{e} = 0.0207 - 0.1033z_1 - 0.2355z_2 + 0.1798z_3$$

The R^2 from this regression is 0.1311, and $NR^2 = 13.11$. The 0.05 critical value for the chi-square distribution with two degrees of freedom is 5.99, so we reject the validity of the two surplus moment conditions. This test does not identify the problem instrument, but since we first tested the validity of z_1 and z_2 and failed to reject their validity, and then found that adding z_3 led us to reject the validity of the surplus moment conditions, the instrument z_3 seems to be the culprit.

10B.2 The Sampling Properties of IV/2SLS

To illustrate the repeated sampling properties of the OLS and IV/2SLS estimators, we use an experimental design based on the discussion in Section 10.4.2. In the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, if x_i is correlated with the error term e_i then x_i is endogenous, and the least squares estimator is biased and inconsistent. An instrumental variable z_i must be correlated with x_i but uncorrelated with e_i in order to be valid. A correlation between z_i and x_i implies that there is a linear association between them. This means that we can describe their relationship as a regression $x_i = \gamma_1 + \theta_1 z_i + v_i$. There is a correlation between x_i and z_i if, and only if, $\theta_1 \neq 0$. If we knew γ_1 and θ_1, we could substitute $E(x_i|z_i) = \gamma_1 + \theta_1 z_i$ into the simple regression model to obtain $y_i = \beta_1 + \beta_2 E(x_i|z_i) + \beta_2 v_i + e_i$. Suppose for a moment that $E(x_i|z_i)$ and v_i can be observed and are viewed as explanatory variables in the regression $y_i = \beta_1 + \beta_2 E(x_i|z_i) + \beta_2 v_i + e_i$. The explanatory variable $E(x_i|z_i)$ is not correlated with the error term e_i because it depends only on z_i. Any correlation between x_i and e_i implies correlation between v_i and e_i because $v_i = x_i - E(x_i|z_i)$.

In the simulation,[15] we use the data generation process $y_i = x_i + e_i$, so that the intercept parameter is 0 and the slope parameter is 1. The first-stage regression is $x_i = \theta z_{i1} + \theta z_{i2} + \theta z_{i3} + v_i$. Note that we have $L = 3$ instruments, each of which has an independent standard normal $N(0,1)$ distribution. The parameter θ controls the instrument strength. If $\theta = 0$, the instruments are not correlated with x_i and instrumental variables estimation will fail. The larger θ becomes the stronger the instruments become. Finally, we create the random errors e_i and v_i to have standard normal distributions with correlation ρ, which controls the endogeneity of x. If $\rho = 0$, then x is not endogenous. The larger ρ becomes the stronger the endogeneity. We create 10,000 samples of size $N = 100$ and then try out OLS and IV/2SLS under several scenarios. We let $\theta = 0.1$ (weak instruments) and $\theta = 0.5$ (strong instruments). We let $\rho = 0$ (x exogenous) and $\rho = 0.8$ (x highly endogenous).

In Table 10B.1, the reported values are

- $\overline{F}$ is the average first-stage F: compare these values to 10. Note that the average value of F is about 2 when $\theta = 0.1$ indicating weak instruments. The average value of F is about 21 when $\theta = 0.5$ indicating strong instruments.

[15] This design is similar to that used by Jinyong Hahn and Jerry Hausman (2003) "Weak Instruments: Diagnosis and Cures in Empirical Economics," *American Economic Review*, 93(2), pp. 118–125.

TABLE 10B.1 **Monte Carlo Simulation Results**

ρ	θ	$\overline{F}$	$\overline{b}_2$	$s.d.(b_2)$	$t(b_2)$	$t(H)$	$\overline{\hat{\beta}}_2$	$s.d.(\hat{\beta}_2)$	$t(\hat{\beta}_2)$
0.0	0.1	1.98	1.0000	0.1000	0.0499	0.0510	0.9941	0.6378	0.0049
0.0	0.5	21.17	0.9999	0.0765	0.0484	0.0518	0.9998	0.1184	0.0441
0.8	0.1	2.00	1.7762	0.0610	1.0000	0.3077	1.3311	0.9483	0.2886
0.8	0.5	21.18	1.4568	0.0610	1.0000	0.9989	1.0111	0.1174	0.0636

- $\overline{b}_2$ is the average of the OLS estimates of $\beta_2 = 1$. The least squares estimator is unbiased when $\rho = 0$, but when $\rho = 0.8$, the least squares estimator shows severe bias.
- $s.d.(b_2)$ is the sample standard deviation of the 10,000 Monte Carlo values of b_2. It tells us how much variation the OLS estimates exhibit in repeated sampling.
- $t(b_2)$ is the percentage of rejections of the true null hypothesis $\beta_2 = 1$ using the 0.05 level of significance t-test based on the OLS estimator. If there is no endogeneity, the percent rejections is very close to the 0.05 value, but if there is strong endogeneity, the OLS estimator rejects the true null hypothesis 100% of the time. That is not good.
- $t(H)$ is the percentage rejections of the regression-based Hausman test for endogeneity using the 0.05 level of significance. If there is no endogeneity, the test rejects 5% of the time, which is what we expect. If there is strong endogeneity but weak instruments, $\theta = 0.1$, the test rejects only 31% of the time, failing to indicate the endogeneity problem. If instruments are not strong, nothing is going to work well. If the instruments are strong, then the test for endogeneity is very successful in detecting strong endogeneity.
- $\overline{\hat{\beta}}_2$ is the average of the instrumental variables estimates of $\beta_2 = 1$. The IV estimator is unbiased when $\rho = 0$. When endogeneity is strong, with weak instruments the IV estimator has a 33% bias, but when instruments are strong it has an average very close to the true value.
- $s.d.(\hat{\beta}_2)$ is the sample standard deviation of the IV estimates in the 10,000 Monte Carlo samples. If there is no endogeneity, note how large its standard deviation is relative to the least squares estimator. With weak instruments its standard deviation is six times that of the least squares estimator. Even with strong instruments, it is substantially larger. The IV estimator is **inefficient** relative to the least squares estimator when endogeneity is absent. When endogeneity is present, the effect of weak instruments shows up in the large standard deviation of the estimates. When instruments are stronger, the standard deviation of the IV estimates falls from 0.95 to 0.12, a substantial improvement.
- Finally, we see the rate of rejections of the true null hypothesis $\beta_2 = 1$ under the scenarios. When x is endogenous and the instruments are weak, the t-test rejects far too often, but it is better than the t-test based on the least squares estimator. Otherwise, the rejection rate is close to the 5% that we expect.

These results are based on a sample size of $N = 100$, which is neither large nor small. What results do you anticipate with larger or smaller samples?

Advice about what to do when there is uncertainty as to whether a regressor is endogenous or not is somewhat mixed. In Table 10.2, the Hausman test statistic p-value is 0.0954. The prevailing attitude is probably summarized by Jeffrey Wooldridge,[16] who says, "We find evidence of endogeneity of *EDUC* at the 10% significance level against a two-sided alternative, and so 2SLS is probably a good idea (assuming that we trust the instruments.)" On the other hand, Patrik

[16] *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition, The MIT Press, 2010, p. 132.

Guggenberger[17] advises, that if testing the coefficient of the endogenous regressor is the objective, then we should avoid considering the Hausman test result and use 2SLS. On the other hand, if we consider how close the estimates are to the true value on average, the "mean square error," Chmelarova and Hill[18] advise that perhaps IV/2SLS should be used only if a Hausman pretest has a much smaller p-value. This result is revealed somewhat in the Monte Carlo simulation. In the case in which $\rho = 0.8$ and $\theta = 0.1$, the mean square error for the least squares estimator is

$$\sum_{m=1}^{10000} \left(b_{2m} - \beta_2\right)^2 / 10000 = 0.6062$$

while for the IV estimator it is

$$\sum_{m=1}^{1000} \left(\hat{\beta}_{2m} - \beta_2\right)^2 / 10000 = 1.0088$$

In other words, in this experimental setting with strong endogeneity and weak instruments, the least squares estimator is, on average, closer to the true parameter value than the IV estimator.

[17] "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test," *Econometric Theory*, 2010, 26(2), pp. 369–382.

[18] "The Hausman Pretest Estimator," *Economics Letters*, 2010, 108, 96–99.

CHAPTER 11

Simultaneous Equations Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain why estimation of a supply and demand model requires an alternative to ordinary least squares (OLS).
2. Explain the difference between exogenous and endogenous variables.
3. Define the ''identification'' problem in simultaneous equations models.
4. Define the reduced form of a simultaneous equations model and explain its usefulness.
5. Explain why it is acceptable to estimate reduced-form equations by least squares.
6. Describe the two-stage least squares estimation procedure for estimating an equation in a simultaneous equations model, and explain how it resolves the estimation problem for least squares.

KEYWORDS

contemporaneous correlation
endogenous variables
exogenous variables
first-stage equation
identification
instrumental variables (IV) estimator
instruments
predetermined variables
reduced-form equation
reduced-form errors
reduced-form parameters
simultaneous equations
structural parameters
two-stage least squares

For most of us, our first encounter with economic models comes through studying supply and demand models, in which the market price and quantity of goods sold are *jointly determined* by the equilibrium of supply and demand. In this chapter, we consider econometric models for data that are jointly determined by two or more economic relations. These **simultaneous equations** models differ from those we have considered in previous chapters because in each model there are *two* or more dependent variables rather than just one.

Simultaneous equations models also differ from most of the econometric models we have considered so far, because they consist of a *set of equations*. For example, price and quantity are determined by the interaction of two equations, one for supply and the other for demand. Simultaneous equations models, which contain more than one dependent variable and more than

one equation, require special statistical treatment. The least squares estimation procedure *is not* appropriate in these models, and we must develop new ways to obtain reliable estimates of economic parameters.

Some of the concepts in this chapter were introduced in Chapter 10. However, reading Chapter 10 is *not* an absolute prerequisite for reading Chapter 11, which is largely self-contained. If you *have* read Chapter 10, you will observe that much of what you learned there will carry over to this chapter, including how simultaneous equations models fit into the big picture. If you *have not* read Chapter 10, referring back to portions of it will provide a deeper understanding of material presented in this chapter. This chapter on simultaneous equations is presented separately because its treatment was the first major contribution of econometrics to the wider field of statistics, and because of its importance in economic analysis.

11.1 A Supply and Demand Model

Supply and demand *jointly* determine the market price of a good and the quantity of it that is sold. Graphically, you recall that market equilibrium occurs at the intersection of the supply and demand curves, as shown in Figure 11.1. An econometric model that explains market price and quantity should consist of two equations, one for supply and the other for demand. It will be a simultaneous equations model, since both equations working together determine price and quantity. A very simple model might look like the following:

$$\text{Demand: } Q_i = \alpha_1 P_i + \alpha_2 X_i + e_{di} \tag{11.1}$$

$$\text{Supply: } Q_i = \beta_1 P_i + e_{si} \tag{11.2}$$

Based on economic theory, we expect the supply curve to be positively sloped, $\beta_1 > 0$, and the demand curve to be negatively sloped, $\alpha_1 < 0$. In this model, we assume that the quantity demanded (Q) is a function of price (P) and income (X). Quantity supplied is taken to be a function of only price. (We have omitted the intercepts to make the algebra easier. In practice, we would include intercept terms in these models.) The observation index $i = 1, \ldots, N$ may represent the market place at different points in time, or at different locations.

The point we wish to make very clear is that it takes *two* equations to describe the supply and demand equilibrium. The *two* equilibrium values, for price and quantity, P^* and Q^*, respectively, are determined at the same time. In this model, the variables P and Q are called **endogenous variables** because their values are determined within the system we have created. The endogenous variables P and Q are *dependent* variables and both are random variables. The income variable X has a value that is determined outside this system. Such variables are said to be **exogenous**, and these variables are treated like usual "x" explanatory variables.

Random errors are added to the supply and demand equations for the usual reasons.

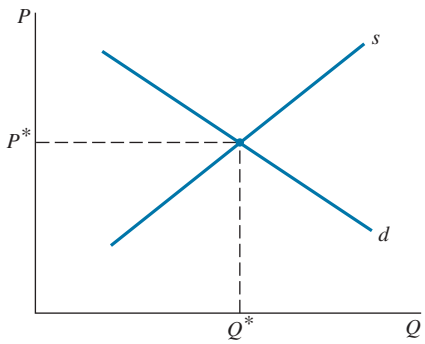


FIGURE 11.1 **Supply and demand equilibrium.**

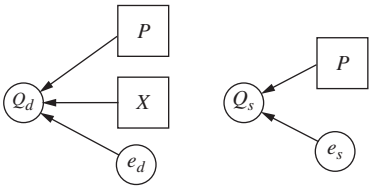


FIGURE 11.2 **Influence diagrams for two regression models.**

We adopt assumption SR2 from Chapter 2 for both the demand and supply equations, given any value of the exogenous variable $X_i, i = 1,\ldots, N$. To simplify notation, we refer to all the values of X_i as $\mathbf{X}$, where $\mathbf{X} = (X_1, X_2, \ldots, X_N)$. Then

$$E(e_{di}|\mathbf{X}) = 0, \qquad E(e_{si}|\mathbf{X}) = 0 \tag{11.3}$$

In Section 2.10, we coined the term "strictly exogenous" for an exogenous variable like this. It implies that $E(e_{di}) = E(e_{si}) = 0$; the unconditional expected value of each error equals zero. It also implies that any value of the exogenous variable X_j is uncorrelated with the error terms in the demand and supply equations, so $\text{cov}(e_{di}, X_j) = 0$ and $\text{cov}(e_{si}, X_j) = 0$. Further, the error terms in the demand and supply equations are assumed to be homoskedastic, $\text{var}(e_{di}|\mathbf{X}) = \sigma_d^2$, and $\text{var}(e_{si}|\mathbf{X}) = \sigma_s^2$. Finally, we also assume that there is no serial correlation and no correlation between the error terms of the two equations.

Let us emphasize the difference between simultaneous equations models and regression models using influence diagrams. An "influence diagram" is a graphical representation of relationships between model components. In the previous chapters, we would have modeled the supply and demand relationships as separate regressions, implying the influence diagrams in Figure 11.2. In this diagram the circles represent endogenous dependent variables and error terms. The squares represent exogenous explanatory variables. In regression analysis, the direction of the influence is one way: from the explanatory variable and the error term to the dependent variable. In this case there is no equilibrating mechanism that will lead quantity demanded to equal quantity supplied at a market-clearing price. For price to adjust to the market-clearing equilibrium, there must be an influence running from P to Q and from Q to P.

Recognizing that price P and quantity Q are *jointly determined*, and that there is feedback between them, suggests the influence diagram in Figure 11.3. In the simultaneous equations model we see the two-way influence, or feedback, between P and Q because they are jointly determined. The random error terms e_d and e_s affect both P and Q, suggesting a correlation between each of the endogenous variables and each of the random error terms. As we will see, this leads to failure of the ordinary least squares (OLS) estimator in simultaneous equations models. Income X is an exogenous variable that affects the endogenous variables, but there is no feedback from P and Q to X.

The fact that P is an endogenous variable on the right-hand side of the supply and demand equations means that we have an explanatory variable that is random. Not only is P random but it is

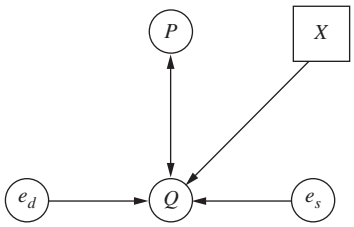


FIGURE 11.3 **Influence diagram for a simultaneous equations model.**

also **contemporaneously correlated** with the random errors in the demand and supply equations, that is, $\text{cov}(P_i, e_{di}) = E(P_i e_{di}) \neq 0$ and $\text{cov}(P_i, e_{si}) = E(P_i e_{si}) \neq 0$. When an explanatory variable is contemporaneously correlated with the regression error term then the OLS estimator is biased and inconsistent. We provide an intuitive argument for why this outcome is true in Section 11.3, and we prove it in Section 11.3.1.

11.2 The Reduced-Form Equations

The two structural equations (11.1) and (11.2) can be solved to express the endogenous variables P and Q as functions of the exogenous variable X. This reformulation of the model is called the **reduced form** of the structural equation system. The reduced form is very important in its own right, and also helps us understand the structural equation system. To find the reduced form, we solve equations (11.1) and (11.2) simultaneously for P and Q.

To solve for P, set Q in the demand and supply equations to be equal,

$$\beta_1 P_i + e_{si} = \alpha_1 P_i + \alpha_2 X_i + e_{di}$$

Then solve for P_i,

$$P_i = \frac{\alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{e_{di} - e_{si}}{(\beta_1 - \alpha_1)} = \pi_1 X_i + v_{1i} \tag{11.4}$$

To solve for Q_i, substitute the value of P_i in (11.4) into either the demand or supply equation. The supply equation is simpler, so substitute P_i into (11.2) and simplify:

$$\begin{aligned} Q_i = \beta_1 P_i + e_{si} &= \beta_1 \left[\frac{\alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{e_{di} - e_{si}}{(\beta_1 - \alpha_1)} \right] + e_{si} \\ &= \frac{\beta_1 \alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{\beta_1 e_{di} - \alpha_1 e_{si}}{(\beta_1 - \alpha_1)} = \pi_2 X_i + v_{2i} \end{aligned} \tag{11.5}$$

The parameters π_1 and π_2 in (11.4) and (11.5) are called **reduced-form parameters**. The errors v_{1i} and v_{2i} are **reduced-form errors**. The reduced forms are predictive equations. We assume that $E(P_i|X_i) = \pi_1 X_i$ and $E(Q_i|X_i) = \pi_2 X_i$. By definition $E(v_{1i}|X_i) = 0$ and $E(v_{2i}|X_i) = 0$, using assumptions (11.3), and also they are homoskedastic and serially uncorrelated if the same holds true for the structural equation errors e_{di} and e_{si}. Under these conditions, the ordinary least squares (OLS) estimators of the reduced-form parameters π_1 and π_2 are consistent, and have approximate normal distributions in large samples, whether the structural equation errors are normal or not. The most important aspect of the OLS estimators for the reduced-form parameters is that they are consistent estimators.

The reduced-form equations (11.4) and (11.5) have an endogenous variable on the left-hand side and exogenous variables, and a random error term, on the right-hand side. These are **first-stage equations** in the language of Chapter 10. We explain the term in Section 11.5 if you have not read Chapter 10. The terms **reduced-form equation** and **first-stage equation** are interchangeable.

The reduced-form equations are important for economic analysis. These equations relate the *equilibrium* values of the endogenous variables to the exogenous variables. Thus, if there is an increase in income X, π_1 is the expected increase in price, after market adjustments lead to a new equilibrium for P and Q. Similarly, π_2 is the expected increase in the expected equilibrium value of Q. (*Question*: how did we determine the directions of these changes?) Secondly, and using the same logic, the estimated reduced-form equations can be used to *predict* values of equilibrium price and quantity for different levels of income. Clearly CEOs and other market analysts are interested in the ability to forecast both prices and quantities sold of their products. Estimating the reduced-form equations makes such predictions possible.

11.3 The Failure of Least Squares Estimation

In this section, we explain why the OLS estimator should not be used to estimate an equation in a simultaneous equations model. For reasons that will become clear in the next section, we focus on the supply equation. In the supply equation (11.2), the endogenous variable P_i on the right-hand side of the equation is *contemporaneously correlated* with the error term e_{si}. Suppose there is a small change, or blip, in the error term e_{si}, say Δe_{si}. Trace the effect of this change through the system. The blip Δe_{si} in the error term of (11.2) is directly transmitted to the equilibrium value of P_i. This follows from the reduced form (11.4) that has P_i on the left and e_{si} on the right. Every change in the supply equation error term, e_{si}, has a direct effect on P_i. Because $\beta_1 > 0$ and $\alpha_1 < 0$, if $\Delta e_{si} > 0$, then $\Delta P_i < 0$. Thus, every time there is a change in e_{si} there is an associated change in P_i in the opposite direction. Consequently, P_i and e_{si} are negatively correlated.

The failure of OLS estimation for the supply equation can be explained as follows: OLS estimation of the relation between Q_i and P_i gives "credit" to price (P_i) for the effect of changes in the error term (e_{si}). This occurs because we do not observe the change in the error term, but only the change in P_i resulting from its correlation with the error e_{si}. The OLS estimator of β_1 will *understate* the true parameter value in this model because of the negative contemporaneous correlation between the endogenous variable P_i and the error term e_{si}. This occurs because we do not observe the change in the error term, but only the change in P_i resulting from its correlation with the error e_{si}. The least squares estimator of β_1 will *understate* the true parameter value in this model because of the negative contemporaneous correlation between the endogenous variable P_i and the error term e_{si}. In large samples, the least squares estimator will tend to be negatively biased in this model. This bias persists even if the sample size goes to infinity, and thus the least squares estimator is inconsistent. This means that the probability distribution of the least squares estimator will ultimately "collapse" about a point that is not the true parameter value as the sample size $N \to \infty$. See Section 5.7 for a general discussion of "large sample" properties of estimators. Here, we summarize by saying:

> The least squares estimator of parameters in a structural simultaneous equation is biased and inconsistent because of the contemporaneous correlation between the random error and the endogenous variables on the right-hand side of the equation.

11.3.1 Proving the Failure of OLS

Consider the supply and demand model in (11.1) and (11.2). To explain the failure of the OLS estimator of the supply equation, let us first obtain the conditional covariance between P_i and e_{si}.

$$
\begin{aligned}
\operatorname{cov}(P_i, e_{si}|\mathbf{X}) &= E\left\{\left[P_i - E(P_i|\mathbf{X})\right]\left[e_{si} - E(e_{si}|\mathbf{X})\right]\middle|\mathbf{X}\right\} \\
&= E(P_i e_{si}|\mathbf{X}) && \left[\text{since } E(e_{si}|\mathbf{X}) = 0\right] \\
&= E\left[(\pi_1 X_i + v_{1i})e_{si}|\mathbf{X}\right] && \left[\text{substitute for } P_i\right] \\
&= E\left[\left(\frac{e_{di} - e_{si}}{\beta_1 - \alpha_1}\right)e_{si}\middle|\mathbf{X}\right] && \left[\text{since } \pi_1 X_i \text{ is fixed}\right] \\
&= \frac{-E(e_{si}^2|\mathbf{X})}{\beta_1 - \alpha_1} && \left[\text{since } e_d, e_s \text{ assumed uncorrelated}\right] \\
&= \frac{-\sigma_s^2}{\beta_1 - \alpha_1} < 0
\end{aligned}
$$

What impact does the negative contemporaneous covariance have on the least squares estimator? The OLS estimator of the supply equation (11.2) (which does not have an intercept term) is

$$b_1 = \frac{\sum P_i Q_i}{\sum P_i^2}$$

Substitute for Q from the reduced-form equation (11.5) and simplify,

$$b_1 = \frac{\sum P_i(\beta_1 P_i + e_{si})}{\sum P_i^2} = \beta_1 + \sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si}$$

The expected value of the least squares estimator is

$$E(b_1|\mathbf{X}) = \beta_1 + E\left[\sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si}\middle|\mathbf{X}\right] = \beta_1 + E\left[\sum\left(\frac{P_i e_{si}}{\sum P_i^2}\right)\middle|\mathbf{X}\right] \quad \text{[move error to numerator]}$$

$$= \beta_1 + \sum\left[E\left(\frac{P_i e_{si}}{\sum P_i^2}\right)\middle|\mathbf{X}\right] \quad \text{[expected value of the sum is sum of expected values]}$$

$$\neq \beta_1 \quad \text{[expected value terms in the sum are not zero]}$$

In the final step, we have $E\left[\left(P_i e_{si}/\sum P_i^2\right)|\mathbf{X}\right] = E\left[g(P_i)e_{si}|\mathbf{X}\right] \neq 0$, where $g(P_i) = P_i/\sum P_i^2$. When finding the covariance between P_i and the random error e_{si}, we showed that $E(P_i e_{si}|\mathbf{X}) = E(P_i e_{si}) = -\sigma_s^2/(\beta_1 - \alpha_1) < 0$ and thus we suspect that $E\left[\left(P_i e_{si}/\sum P_i^2\right)|\mathbf{X}\right] < 0$, because $\sum P_i^2 > 0$, so that we suspect the least squares estimator exhibits a negative bias. However, the expected value of the ratio is not the ratio of expected values, so all we can really conclude is that the least squares estimator is biased, because e_{si} and P_i are contemporaneously correlated.

This bias does not disappear in larger samples, so the OLS estimator of the supply equation is inconsistent as well. The OLS estimator converges to a value less than β_1 and this is easier to show using asymptotic analysis similar to that in Chapter 5, equation (5.41). Rewrite the OLS estimators

$$b_1 = \beta_1 + \sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si} = \beta_1 + \frac{\sum P_i e_{si}}{\sum P_i^2} = \beta_1 + \frac{\sum P_i e_{si}/N}{\sum P_i^2/N} = \beta_1 + \frac{\widehat{E(P_i e_{si})}}{\widehat{E(P_i^2)}}$$

Using the Law of Large Numbers, sample moments (averages) converge to population moments (expected values), so that

$$\widehat{E(P_i e_{si})} \xrightarrow{p} E(P_i e_{si}) = -\sigma_s^2/(\beta_1 - \alpha_1) < 0$$

and

$$\widehat{E(P_i^2)} \xrightarrow{p} E(P_i^2) > 0$$

Therefore

$$b_1 \xrightarrow{p} \beta_1 - \frac{\sigma_s^2/(\beta_1 - \alpha_1)}{E(P_i^2)} < \beta_1$$

11.4 The Identification Problem

In the supply and demand model given by (11.1) and (11.2),

- The parameters of the demand equation, α_1 and α_2, *cannot* be consistently estimated by *any* estimation method.
- The slope of the supply equation, β_1, *can* be consistently estimated.

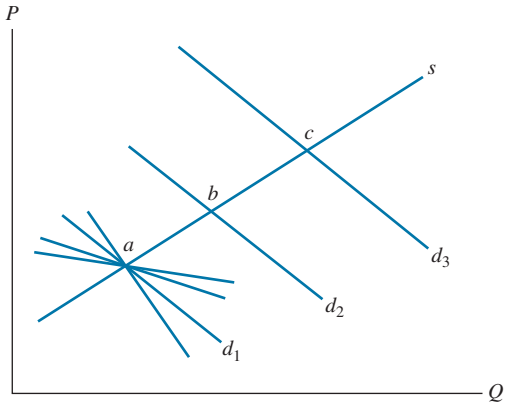


FIGURE 11.4 **The effect of changing income.**

How are we able to make such statements? The answer is quite intuitive, and it can be illustrated graphically. What happens when income X changes? The demand curve shifts and a new equilibrium price and quantity are created. In Figure 11.4 we show the demand curves d_1, d_2, and d_3 and equilibria, at points a, b, and c, for three levels of income. As income changes, data on price and quantity will be observed around the intersections of supply and demand. The random errors e_d and e_s cause small shifts in the supply and demand curves, creating equilibrium observations on price and quantity that are scattered about the intersections at points a, b, and c.

The data values will trace out the *supply curve*, suggesting that we can fit a line through them to estimate the slope β_1. The data values fall along the supply curve because income is *present* in the demand curve and *absent* from the supply curve. As income changes, the demand curve shifts but the supply curve remains fixed, resulting in observations along the supply curve.

There are *no* data values falling along any of the demand curves, and there is no way to estimate their slope. Any one of the infinite number of demand curves passing through the equilibrium points could be correct. Given the data, there is no way to distinguish the true demand curve from all the rest. Through the equilibrium point a we have drawn a few demand curves, each of which *could* have generated the data we observe.

The problem lies with the model that we are using. There is no variable in the supply equation that will shift it relative to the demand curve. If we were to add a variable to the supply curve, say W, then each time W is changed, the supply curve would shift, and the demand curve would stay fixed. The shifting of supply relative to a fixed demand curve (since W is *absent* from the demand equation) would create equilibrium observations along the demand curve, making it possible to estimate the slope of the demand curve and the effect of income on demand.

It is the *absence* of variables in one equation that are *present* in another equation that makes parameter estimation possible. A general rule, which is called a **necessary condition** for **identification** of an equation, is this:

A Necessary Condition for Identification

In a system of M simultaneous equations, which jointly determine the values of M endogenous variables, at least $M - 1$ variables must be absent from an equation for estimation of its parameters to be possible. When estimation of an equation's parameters is possible, then the equation is said to be *identified*, and its parameters can be estimated consistently. If fewer than $M - 1$ variables are omitted from an equation, then it is said to be *unidentified*, and its parameters cannot be consistently estimated.

In our supply and demand model there are $M = 2$ equations, so we require at least $M - 1 = 1$ variable to be omitted from an equation to identify it. There are a total of three variables: P, Q, and X. In the demand equation none of the variables are omitted; thus it is unidentified and its parameters cannot be estimated consistently. In the supply equation, one variable, income (X), is omitted; the supply curve is identified, and its parameter can be estimated.

The identification condition must be checked *before* trying to estimate an equation. If an equation is not identified, then changing the model must be considered before it is estimated. However, changing the model should not be done in a haphazard way; no important variable should be omitted from an equation just to identify it. The structure of a simultaneous equations model should reflect your understanding of how equilibrium is achieved and should be consistent with economic theory. Creating a false model is not a good solution to the identification problem.

This paragraph is for those who have read Chapter 10. The necessary condition for identification can be expressed in an alternative but equivalent fashion. The two-stage least squares estimation procedure was developed in Chapter 10 and shown to be an **instrumental variables estimator**. This procedure is developed further in the next section. The number of instrumental variables (IVs) required for estimation of an equation within a simultaneous equations model is equal to the number of right-hand side endogenous variables. In a typical equation within a simultaneous equations model, several **exogenous variables** appear on the right-hand side. Thus **instruments** must come from those exogenous variables omitted from the equation in question. Consequently, identification requires that the number of excluded exogenous variables in an equation be at least as large as the number of included right-hand side endogenous variables. This ensures an adequate number of IVs.

11.5 Two-Stage Least Squares Estimation

The most widely used method for estimating the parameters of an identified structural equation is called **two-stage least squares**, which is often abbreviated as 2SLS or TSLS. The name comes from the fact that it can be calculated using two OLS regressions. We will explain how it works by considering the supply equation in (11.2). Recall that we should not apply the usual OLS procedure to estimate β_1 in this equation because the endogenous variable P_i on the right-hand side of the equation is contemporaneously correlated with the error term e_{si}, causing the OLS estimator to be biased and inconsistent.

The variable P_i is composed of a systematic part, which is its expected value $E(P_i|X_i)$, and a random part, which is the reduced-form random error v_{1i}, that is,

$$P_i = E(P_i|X_i) + v_{1i} \tag{11.6}$$

In the supply equation (11.2), the portion of P_i that causes problems for the OLS estimator is v_{1i}, the random part. It is v_{1i} that causes P_i to be correlated with the error term e_{si}. If we knew $E(P_i|X_i)$, then we could replace P_i in (11.2) by (11.6) to obtain

$$Q_i = \beta_1 [E(P_i|X_i) + v_{1i}] + e_{si} = \beta_1 E(P_i|X_i) + (\beta_1 v_{1i} + e_{si}) \tag{11.7}$$

In (11.7) the explanatory variable on the right-hand side is $E(P_i|X_i)$. It depends only on the exogenous variable, and it is not correlated with the error term. We could apply OLS to (11.7) to consistently estimate β_1.

Of course, we cannot use the variable $E(P_i|X_i)$ in place of P_i since we do not know it. However, we can consistently estimate $E(P_i|X_i)$. Let $\hat{\pi}_1$ come from the fitted OLS estimation of the reduced-form equation for P_i. A consistent estimator for $E(P_i|X_i)$ is

$$\hat{P}_i = \hat{\pi}_1 X_i$$

Using $\hat{P}_i$ as a replacement for $E(P_i|X_i)$ in (11.7), we obtain

$$Q_i = \beta_1 \hat{P}_i + \hat{e}_{*i} \tag{11.8}$$

In large samples, $\hat{P}_i$ and the random error $\hat{e}_{*i}$ are uncorrelated, and consequently the parameter β_1 can be consistently estimated by applying OLS to (11.8).

The OLS estimator of (11.8) is the **two-stage least squares** estimator of β_1, which is consistent and asymptotically normal. Because the two-stage least squares estimator is consistent it converges to the true value in large samples. That the estimator is asymptotically normal means that if we have a large sample, the usual tests and confidence interval estimators can be used. To summarize, the two stages of the estimation procedure are:

1. OLS estimation of the reduced-form equation for P_i and the calculation of its predicted value, $\hat{P}_i$
2. OLS estimation of the structural equation in which the right-hand side endogenous variable P_i is replaced by its predicted value $\hat{P}_i$[1]

In practice always use software that is designed for 2SLS, so that standard errors and t-values will be calculated correctly.

11.5.1 The General Two-Stage Least Squares Estimation Procedure

The two-stage least squares estimation procedure can be used to estimate the parameters of any identified equation within a simultaneous equations system. In a system of M simultaneous equations, let the endogenous variables be y_{i1}, y_{i2},…, y_{iM}. There must always be as many equations in a simultaneous system as there are endogenous variables. Let there be K exogenous variables, $x_{i1}, x_{i2},\ldots, x_{iK}$. To illustrate, suppose $M = 3$ and the first structural equation within this system is

$$y_{i1} = \alpha_2 y_{i2} + \alpha_3 y_{i3} + \beta_1 x_{i1} + \beta_2 x_{i2} + e_{i1} \tag{11.9}$$

If this equation is identified, then its parameters can be estimated in two steps:

1. Use OLS to estimate the parameters of the reduced-form equations

$$y_{i2} = \pi_{12} x_{i1} + \pi_{22} x_{i2} + \cdots + \pi_{K2} x_{iK} + v_{i2}$$

$$y_{i3} = \pi_{13} x_{i1} + \pi_{23} x_{i2} + \cdots + \pi_{K3} x_{iK} + v_{i3}$$

   Obtain the predicted values

$$\hat{y}_{i2} = \hat{\pi}_{12} x_{i1} + \hat{\pi}_{22} x_{i2} + \cdots + \hat{\pi}_{K2} x_{iK}$$

$$\hat{y}_{i3} = \hat{\pi}_{13} x_{i1} + \hat{\pi}_{23} x_{i2} + \cdots + \hat{\pi}_{K3} x_{iK} \tag{11.10}$$

2. Replace the endogenous variables, y_{i2} and y_{i3}, on the right-hand side of the structural (11.9) by their predicted values from (11.10)

$$y_{i1} = \alpha_2 \hat{y}_{i2} + \alpha_3 \hat{y}_{i3} + \beta_1 x_{i1} + \beta_2 x_{i2} + e^*_{i1}$$

   Estimate the parameters of this equation by OLS.

In practice, we should always use software designed for 2SLS or IV estimation. It will correctly carry out the calculations of the 2SLS estimates and their standard errors.

Equation (11.9) has two right-hand side endogenous variables and two exogenous variables. K is the total number of exogenous variables. How large must K be so that equation (11.9) is identified? The identification "necessary" condition is that in a system of M equations at

[1]The discussion above is an intuitive explanation of the two-stage least squares estimator. For a general explanation of this estimation method, see Section 10.3. There we derive the two-stage least squares estimator and discuss its properties.

least $M - 1$ variables that appear elsewhere in the system must be omitted from each equation. There are $M = 3$ equations so $M - 1 = 2$ variables must be omitted from each equation. Let $K = K_1 + K_1^*$, where $K_1 = 2$ is the number of included exogenous variables in the first structural equation, and K_1^* is the number of exogenous variables excluded from the first structural equation. Identification of the first equation requires $K_1^* \geq 2$ and $K \geq 4$. In Chapter 10's terminology, K_1^* is the number of instrumental variables for the first equation.

The alternative description of the condition for identification is that the number of omitted exogenous variables, K_1^*, must be greater than, or equal to, the number of included, right-hand side endogenous variables. Let $M = 1 + M_1 + M_1^*$, where $M_1 = 2$ is the number of included right-hand side endogenous variables, and M_1^* is the number of endogenous variables excluded from the first equation. In this example, $M_1^* = 0$ because the first equation contains all three endogenous variables, including the left-hand side variable y_1. The identification rule is that $K_1^* \geq M_1$. In Chapter 10's language, there must be as many instrumental variables, K_1^*, as endogenous variables on the right-hand side of the equation, M_1.

Remark

Simultaneous equations models were developed in the early 1940s and for many years were the cornerstone of econometric analysis. The subject of Chapter 10 is regression equations with endogenous variables, which can be thought of as one equation from a system of equations. Because building and estimating complete systems are difficult, more researchers in recent years have relied on estimating individual equations by $2SLS/IV$, which is why the content of Chapter 10 precedes this treatment of simultaneous equations. However, the concepts and methods used in Chapters 10 and 11 are the same. Just keep in mind that:

1. **Two-stage least squares** and **instrumental variables estimation** are identical.
2. **IVs**, or just **instruments**, are exogenous variables that do not appear in the equation. Instruments are **excluded exogenous variables**.
3. **The reduced-form equations** in simultaneous equations modeling are the **first-stage equations** in instrumental variables, two-stage least squares, estimation.

11.5.2 The Properties of the Two-Stage Least Squares Estimator

We have described how to obtain estimates for structural equation parameters in identified equations. The properties of the two-stage least squares estimator are as follows:

- The 2SLS estimator is a biased estimator, but it is consistent.
- In large samples the 2SLS estimator is approximately normally distributed.
- The variances and covariances of the 2SLS estimator are unknown in small samples, but for large samples, we have expressions for them that we can use as approximations. These formulas are built into econometric software packages, which report standard errors and t-values, just like an OLS regression program.
- If you obtain 2SLS estimates by applying two least squares regressions using OLS regression software, the standard errors and t-values reported in the *second* regression are *not* correct for the 2SLS estimator. Always use specialized 2SLS or IV software when obtaining estimates of structural equations.

EXAMPLE 11.1 | Supply and Demand for Truffles

Truffles are a gourmet delight. They are edible fungi that grow below the ground. In France they are often located by collectors who use pigs to sniff out the truffles and "point" to them. Actually the pigs dig frantically for the truffles because pigs have an insatiable taste for them, as do the French, and they must be restrained from "pigging out" on them. Consider a supply and demand model for truffles:

$$\text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 PS_i + \alpha_4 DI_i + e_{di} \quad (11.11)$$

$$\text{Supply: } Q_i = \beta_1 + \beta_2 P_i + \beta_3 PF_i + e_{si} \quad (11.12)$$

In the demand equation Q is the quantity of truffles traded in a particular French marketplace, indexed by i, P is the market price of truffles, PS is the market price of a substitute for real truffles (another fungus much less highly prized), and DI is per capita monthly disposable income of local residents. The supply equation contains the market price and quantity supplied. Also it includes PF, the price of a factor of production, which in this case is the hourly rental price of truffle-pigs used in the search process. In this model, we assume that P and Q are endogenous variables. The exogenous variables are PS, DI, PF, and the intercept.

Identification

Before thinking about estimation, check the identification of each equation. The rule for identifying an equation is that in a system of M equations at least $M-1$ variables must be omitted from each equation in order for it to be identified. In the demand equation the variable PF is not included; thus the necessary $M-1=1$ variable is omitted. In the supply equation both PS and DI are absent; more than enough to satisfy the identification condition. Note too that the variables that are omitted are different for each equation, ensuring that each contains at least one *shift* variable not present in the other. We conclude that each equation in this system is identified and can thus be estimated by two-stage least squares.

Why are the variables omitted from their respective equations? Because economic theory says that the price of a factor of production should affect supply but not demand, and that the price of substitute goods and income should affect demand and not supply. The specifications we used are based on the microeconomic theory of supply and demand.

The reduced-form equations

The reduced-form equations express each endogenous variable, P and Q, in terms of the exogenous variables PS, DI, PF, and the intercept, plus an error term. They are

$$Q_i = \pi_{11} + \pi_{21} PS_i + \pi_{31} DI_i + \pi_{41} PF_i + v_{i1}$$

$$P_i = \pi_{12} + \pi_{22} PS_i + \pi_{32} DI_i + \pi_{42} PF_i + v_{i2}$$

We can estimate these equations by OLS since the right-hand side variables are exogenous and contemporaneously uncorrelated with the random errors v_{i1} and v_{i2}. The data file *truffles* contains 30 observations on each of the endogenous and exogenous variables. The units of measurement are \$ per ounce for price P, ounces for Q, \$ per ounce for PS, and thousands of dollars for DI; PF is the hourly rental rate (\$) for a truffle-finding pig. A few of the observations are shown in Table 11.1. The results of the least squares estimations of the reduced-form equations for Q and P are reported in Tables 11.2a and 11.2b.

In Table 11.2a, we see that the estimated coefficients are statistically significant, and thus we conclude that the exogenous variables affect the quantity of truffles traded, Q, in this reduced-form equation. The $R^2 = 0.697$, and the overall F-statistic is 19.973, which has a p-value of less than 0.0001. In Table 11.2b the estimated coefficients

TABLE 11.1 Representative Truffle Data

OBS	P	Q	PS	DI	PF
1	29.64	19.89	19.97	2.103	10.52
2	40.23	13.04	18.04	2.043	19.67
3	34.71	19.61	22.36	1.870	13.74
4	41.43	17.13	20.87	1.525	17.95
5	53.37	22.55	19.79	2.709	13.71
Summary Statistics					
Mean	62.72	18.46	22.02	3.53	22.75
Std. Dev.	18.72	4.61	4.08	1.04	5.33

TABLE 11.2a Reduced Form for Quantity of Truffles (Q)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.8951	3.2434	2.4342	0.0221
PS	0.6564	0.1425	4.6051	0.0001
DI	2.1672	0.7005	3.0938	0.0047
PF	−0.5070	0.1213	−4.1809	0.0003

TABLE 11.2b Reduced Form for Price of Truffles (P)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−32.5124	7.9842	−4.0721	0.0004
PS	1.7081	0.3509	4.8682	0.0000
DI	7.6025	1.7243	4.4089	0.0002
PF	1.3539	0.2985	4.5356	0.0001

are statistically significant, indicating that the exogenous variables have an effect on market price P. The $R^2 = 0.889$ implies a good fit of the reduced-form equation to the data. The overall F-statistic value is 69.189 that has a p-value of less than 0.0001, indicating that the model has statistically significant explanatory power.

The structural equations

The reduced-form equations are used to obtain $\hat{P}$ that will be used in place of P on the right-hand side of the supply and demand equations in the second stage of two-stage least squares. From Table 11.2b, we have

$$\hat{P}_i = \hat{\pi}_{12} + \hat{\pi}_{22}PS_i + \hat{\pi}_{32}DI_i + \hat{\pi}_{42}PF_i$$
$$= -32.512 + 1.708PS_i + 7.603DI_i + 1.354PF_i$$

The 2SLS results are given in Tables 11.3a and 11.3b. The estimated demand curve results are in Table 11.3a. Note that the coefficient of price is negative, indicating that as the market price rises, the quantity demanded of truffles declines, as predicted by the law of demand. The standard errors that are reported are obtained from 2SLS software. They and the t-values are valid in large samples. The p-value indicates that the estimated slope of the demand curve is significantly different from zero. Increases in the price of the substitute for truffles increase the demand for truffles, which is a characteristic of substitute goods. Finally the effect of income is positive, indicating that truffles are a normal good. All of these variables have statistically significant coefficients and thus have an effect upon the quantity demanded.

TABLE 11.3a **2SLS Estimates for Truffle Demand**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−4.2795	5.5439	−0.7719	0.4471
P	−0.3745	0.1648	−2.2729	0.0315
PS	1.2960	0.3552	3.6488	0.0012
DI	5.0140	2.2836	2.1957	0.0372

TABLE 11.3b **2SLS Estimates for Truffle Supply**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.0328	1.2231	16.3785	0.0000
P	0.3380	0.0249	13.5629	0.0000
PF	−1.0009	0.0825	−12.1281	0.0000

The supply equation results appear in Table 11.3b. As anticipated, increases in the price of truffles increase the quantity supplied, and increases in the rental rate for truffle-seeking pigs, which is an increase in the cost of a factor of production, reduces supply. Both of these variables have statistically significant coefficient estimates.

EXAMPLE 11.2 | Supply and Demand at the Fulton Fish Market

The Fulton Fish Market has operated in New York City for over 150 years. The prices for fish are determined daily by the forces of supply and demand. Kathryn Graddy[2] collected daily data on the price of whiting (a common type of fish), quantities sold, and weather conditions during the period December 2, 1991, to May 8, 1992. These data are in the file *fultonfish*. Fresh fish arrive at the market about midnight. The wholesalers, or dealers, sell to buyers for retail shops and restaurants. The first interesting feature of this example is to consider whether prices and quantities are *simultaneously* determined by supply and demand at all.[3] We might consider this a market with a fixed, perfectly inelastic supply. At the start of the day, when the market is opened, the supply of fish available for the day is fixed. If supply is fixed, with a vertical supply curve, then price is demand-determined, with higher demand leading to higher prices but no increase in the quantity supplied. If this is true, then the feedback between prices and quantities is eliminated. Such models are said to be **recursive** and the demand equation can be estimated by OLS rather than the more complicated two-stage least squares procedure.

However whiting fish can be kept for several days before going bad, and dealers can decide to sell less, and add to their inventory, or buffer stock, if the price is judged too low, in hope for better prices the next day. Or, if the price is unusually high on a given day, then sellers can increase the day's catch with additional fish from their buffer stock. Thus despite the perishable nature of the product, and the daily resupply of fresh fish, daily price is simultaneously determined by supply and demand forces. The key point here is that "simultaneity" does not require that events occur at a simultaneous moment in time.

[2] See Kathryn Graddy (2006), "The Fulton Fish Market," *Journal of Economic Perspectives*, 20(2), 207–220.

[3] See Kathryn Graddy and Peter E. Kennedy (2010), "When Are Supply and Demand Determined Recursively Rather than Simultaneously?," *Eastern Economic Journal*, 36, 188–197.

Let us specify the demand equation for this market as

$$\ln(QUAN_t) = \alpha_1 + \alpha_2 \ln(PRICE_t) + \alpha_3 MON_t + \alpha_4 TUE_t + \alpha_5 WED_t + \alpha_6 THU_t + e_{dt} \quad (11.13)$$

where $QUAN_t$ is the quantity sold, in pounds, and $PRICE_t$ is the average daily price per pound. Note that we are using the subscript "t" to index observations for this relationship because of the time series nature of the data. The remaining variables are indicator variables for the days of the week, with Friday being omitted. The coefficient α_2 is the price elasticity of demand, which we expect to be negative. The daily indicator variables capture day-to-day shifts in demand. The supply equation is

$$\ln(QUAN_t) = \beta_1 + \beta_2 \ln(PRICE_t) + \beta_3 STORMY_t + e_{st} \quad (11.14)$$

The coefficient β_2 is the price elasticity of supply. The variable *STORMY* is an indicator variable indicating stormy weather during the previous three days. This variable is important in the supply equation because stormy weather makes fishing more difficult, reducing the supply of fish brought to market.

Identification

Prior to estimation, we must determine whether the supply and demand equation parameters are identified. The necessary condition for an equation to be identified is that in this system of $M = 2$ equations, it must be true that at least $M - 1 = 1$ variable must be omitted from each equation. In the demand equation the weather variable *STORMY* is omitted, and it does appear in the supply equation. In the supply equation, the four daily indicator variables that are included in the demand equation are omitted. Thus the demand equation shifts daily, while the supply remains fixed (since the supply equation does not contain the daily indicator variables), thus tracing out the supply curve, making it identified, as shown in Figure 11.4. Similarly, stormy conditions shift the supply curve relative to a fixed demand, tracing out the demand curve and making it identified.

The reduced-form equations

The reduced-form equations specify each endogenous variable as a function of all exogenous variables

$$\ln(QUAN_t) = \pi_{11} + \pi_{21} MON_t + \pi_{31} TUE_t + \pi_{41} WED_t + \pi_{51} THU_t + \pi_{61} STORMY_t + v_{t1} \quad (11.15)$$

$$\ln(PRICE_t) = \pi_{12} + \pi_{22} MON_t + \pi_{32} TUE_t + \pi_{42} WED_t + \pi_{52} THU_t + \pi_{62} STORMY_t + v_{t2} \quad (11.16)$$

These reduced-form equations can be estimated by OLS because the right-hand side variables are all exogenous and uncorrelated with the reduced-form errors v_{t1} and v_{t2}. Using the Graddys' data (*fultonfish*), we estimate these reduced-form equations and report them in Tables 11.4a and 11.4b. Estimation of the reduced-form equations is the first step of two-stage least squares estimation of the supply and demand equations. It is a requirement for successful two-stage least squares estimation that the estimated coefficients in the reduced form for the right-hand side endogenous variable be statistically significant. We have specified the structural equations (11.13) and (11.14) with $\ln(QUAN_t)$ as the left-hand side variable and $\ln(PRICE_t)$ as the right-hand side endogenous variable. Thus the key reduced-form equation is (11.16) for $\ln(PRICE_t)$. In this equation

- To identify the supply curve, the daily indicator variables must be jointly significant. This implies that at least one of their coefficients is statistically different from zero, meaning that there is at least one significant shift variable in the demand equation, which permits us to reliably estimate the supply equation.
- To identify the demand curve, the variable $STORMY_t$ must be statistically significant, meaning that supply has a significant shift variable, so that we can reliably estimate the demand equation.

Why is this so? The identification discussion in Section 11.4 requires only the presence of shift variables, not their significance. The answer comes from a great deal of econometric research in the past decade, which shows that the two-stage least squares estimator performs very poorly if the shift variables are not strongly significant.[4] Recall that to implement two-stage least squares we take the predicted value from the reduced-form regression and include it in the structural equations in place of the right-hand side endogenous variable, that is, we calculate

$$\widehat{\ln(PRICE_t)} = \hat{\pi}_{12} + \hat{\pi}_{22} MON_t + \hat{\pi}_{32} TUE_t + \hat{\pi}_{42} WED_t + \hat{\pi}_{52} THU_t + \hat{\pi}_{62} STORMY_t$$

where $\hat{\pi}_{k2}$ are the least squares estimates of the reduced-form coefficients, and then replace $\ln(PRICE_t)$ with $\widehat{\ln(PRICE_t)}$. To illustrate our point, let us focus on the problem of estimating the supply equation (11.14) and take the extreme case that $\hat{\pi}_{22} = \hat{\pi}_{32} = \hat{\pi}_{42} = \hat{\pi}_{52} = 0$, meaning that the coefficients on the daily indicator variables are all identically zero. Then

$$\widehat{\ln(PRICE_t)} = \hat{\pi}_{12} + \hat{\pi}_{62} STORMY_t$$

If we replace $\ln(PRICE_t)$ in the supply equation (11.14) with this predicted value, there will be *exact* collinearity between $\widehat{\ln(PRICE_t)}$ and the variable $STORMY_t$, which is already in the supply equation, and two-stage least squares will fail. If the coefficient estimates on the daily indicator

[4] See Section 10.3.9 for further discussion of this point.

variables are not exactly zero, but are jointly insignificant, it means there will be severe collinearity in the second stage, and although the two-stage least squares estimates of the supply equation can be computed, they will be unreliable. In Table 11.4b, showing the reduced-form estimates for (11.16), none of the daily indicator variables are statistically significant. Also, the joint F-test of significance of the daily indicator variables has p-value 0.65, so that we cannot reject the null hypothesis that all these coefficients are zero.[5] In this case the supply equation is not identified in practice, and we will not report estimates for it.

TABLE 11.4a Reduced Form for ln(Quantity) Fish

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.8101	0.1470	59.9225	0.0000
STORMY	−0.3878	0.1437	−2.6979	0.0081
MON	0.1010	0.2065	0.4891	0.6258
TUE	−0.4847	0.2011	−2.4097	0.0177
WED	−0.5531	0.2058	−2.6876	0.0084
THU	0.0537	0.2010	0.2671	0.7899

TABLE 11.4b Reduced Form for ln(Price) Fish

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−0.2717	0.0764	−3.5569	0.0006
STORMY	0.3464	0.0747	4.6387	0.0000
MON	−0.1129	0.1073	−1.0525	0.2950
TUE	−0.0411	0.1045	−0.3937	0.6946
WED	−0.0118	0.1069	−0.1106	0.9122
THU	0.0496	0.1045	0.4753	0.6356

However, $STORMY_t$ is statistically significant in Table 11.4b, meaning that the demand equation may be reliably estimated by two-stage least squares. An advantage of two-stage least squares estimation is that each equation can be treated and estimated separately, so the fact that the supply equation is not reliably estimable does not mean that we cannot proceed with estimation of the demand equation. The check of statistical significance of the sets of shift variables for the structural equations should be carried out each time a simultaneous equations model is formulated.

Two-stage least squares estimation of fish demand

Applying two-stage least squares estimation to the demand equation we obtain the results as given in Table 11.5. The price elasticity of demand is estimated to be −1.12, meaning that a 1% increase in fish price leads to about a 1.12% decrease in the quantity demanded; this estimate is statistically significant at the 5% level. The indicator variable coefficients are negative and statistically significant for Tuesday and Wednesday, meaning that demand is lower on these days relative to Friday.

TABLE 11.5 2SLS Estimates for Fish Demand

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.5059	0.1662	51.1890	0.0000
ln(*PRICE*)	−1.1194	0.4286	−2.6115	0.0103
MON	−0.0254	0.2148	−0.1183	0.9061
TUE	−0.5308	0.2080	−2.5518	0.0122
WED	−0.5664	0.2128	−2.6620	0.0090
THU	0.1093	0.2088	0.5233	0.6018

EXAMPLE 11.3 | Klein's Model I

One of the most widely used econometric examples in the past 50 years is the small, three equation, macroeconomic model of the U.S. economy proposed by Lawrence Klein, the 1980 Nobel Prize winner in Economics.[6] The model has three equations, which are estimated, and then a number of macroeconomic identities, or definitions, to complete the model. In all, there are eight endogenous variables and eight exogenous variables.

[5] Even if the variables are jointly significant, there may be a problem. The significance must be "strong." An F-value <10 is cause for concern. This problem is the same as that of weak instruments in instrumental variables estimation (see Section 10.3.9).

[6] Our presentation follows Ernst R. Berndt (1991), *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley Publishing, Section 10.5.

The first equation is a consumption function, in which aggregate consumption in year t, CN_t is related to total wages earned by all workers, W_t. Total wages are divided into wages of workers earned in the private sector, W_{1t}, and wages of workers earned in the public sector, W_{2t}, so that total wages $W_t = W_{1t} + W_{2t}$. Private sector wages W_{1t} are endogenous and determined within the structure of the model, as we will see below. Public sector wages W_{2t} are exogenous. In addition, consumption expenditures are related to nonwage income (profits) in the current year, P_t, which are endogenous, and profits from the previous year, P_{t-1}. Thus, the consumption function is

$$CN_t = \alpha_1 + \alpha_2\left(W_{1t} + W_{2t}\right) + \alpha_3 P_t + \alpha_4 P_{t-1} + e_{1t} \quad (11.17)$$

Now refer back to equation (5.44) in Section 5.7.3. There we introduced the term **contemporaneously uncorrelated** to describe the situation in which an explanatory variable observed at time t, x_{tk} is uncorrelated with the random error at time t, e_t. In the terminology of Chapter 10, the variable x_{tk} is **exogenous** if it is contemporaneously uncorrelated with the random error e_t. And the variable x_{tk} is **endogenous** if it is contemporaneously correlated with the random error e_t. In the consumption equation, W_{1t} and P_t are endogenous and contemporaneously correlated with the random error e_t. On the other hand, wages in the public sector, W_{2t}, are set by public authority and are assumed exogenous and uncorrelated with the current period random error e_{1t}. What about profits in the previous year, P_{t-1}? They are **not** correlated with the random error occurring one year later. Lagged endogenous variables are called **predetermined variables** and are treated just like exogenous variables.

The second equation in the model is the investment equation. Net investment, I_t, is specified to be a function of current and lagged profits, P_t and P_{t-1}, as well as the capital stock at the end of the previous year, K_{t-1}. This lagged variable is predetermined and treated as exogenous. The investment equation is

$$I_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + e_{2t} \quad (11.18)$$

Finally, there is an equation for wages in the private sector, W_{1t}. Let $E_t = CN_t + I_t + \left(G_t - W_{2t}\right)$, where G_t is government spending. Consumption and investment are endogenous and government spending and public sector wages are exogenous. The sum, E_t, total national product minus public sector wages, is endogenous. Wages are taken to be related to E_t and the predetermined variable E_{t-1}, plus a time trend variable, $TIME_t = YEAR_t - 1931$, which is exogenous. The wage equation is

$$W_{1t} = \gamma_1 + \gamma_2 E_t + \gamma_3 E_{t-1} + \gamma_4 TIME_t + e_{3t} \quad (11.19)$$

Because there are eight endogenous variables in the entire system there must also be eight equations. Any system of M endogenous variables must have M equations to be complete. In addition to the three equations (11.17)–(11.19), which contain five endogenous variables, there are five other definitional equations to complete the system that introduce three further endogenous variables. In total, there are eight exogenous and predetermined variables, which can be used as IVs. The exogenous variables are government spending, G_t, public sector wages, W_{2t}, taxes, TX_t, and the time trend variable, $TIME_t$. Another exogenous variable is the constant term, the "intercept" variable in each equation, $X_{1t} \equiv 1$. The predetermined variables are lagged profits, P_{t-1}, the lagged capital stock, K_{t-1}, and the lagged total national product minus public sector wages, E_{t-1}.

11.6 Exercises

11.6.1 Problems

11.1 Our aim is to estimate the parameters of the simultaneous equations model

$$y_1 = \alpha_1 y_2 + e_1$$
$$y_2 = \alpha_2 y_1 + \beta_1 x_1 + \beta_2 x_2 + e_2$$

We assume that x_1 and x_2 are exogenous and uncorrelated with the error terms e_1 and e_2.

a. Solve the two structural equations for the reduced-form equation for y_2, that is, $y_2 = \pi_1 x_1 + \pi_2 x_2 + v_2$. Express the reduced-form parameters in terms of the **structural parameters** and the reduced-form error in terms of the structural parameters and e_1 and e_2. Show that y_2 is correlated with e_1.

b. Which equation parameters are consistently estimated using OLS? Explain.

c. Which parameters are "identified," in the simultaneous equations sense? Explain your reasoning.

d. To estimate the parameters of the reduced-form equation for y_2 using the method of moments (MOM), which was introduced in Section 10.3, the two moment equations are

$$N^{-1}\sum x_{i1}(y_2 - \pi_1 x_{i1} - \pi_2 x_{i2}) = 0$$
$$N^{-1}\sum x_{i2}(y_2 - \pi_1 x_{i1} - \pi_2 x_{i2}) = 0$$

Explain why these two moment conditions are a valid basis for obtaining consistent estimators of the reduced-form parameters.

e. Are the MOM estimators in part (d) the same as the OLS estimators? Form the sum of squared errors function for $y_2 = \pi_1 x_1 + \pi_2 x_2 + v_2$ and find the first derivatives. Set these to zero and show that they are equivalent to the two equations in part (d).

f. Using $\sum x_{i1}^2 = 1$, $\sum x_{i2}^2 = 1$, $\sum x_{i1}x_{i2} = 0$, $\sum x_{i1}y_{1i} = 2$, $\sum x_{i1}y_{2i} = 3$, $\sum x_{i2}y_{1i} = 3$, $\sum x_{i2}y_{2i} = 4$, and the two moment conditions in part (d) show that the MOM/OLS estimates of π_1 and π_2 are $\hat{\pi}_1 = 3$ and $\hat{\pi}_2 = 4$.

g. The fitted value $\hat{y}_2 = \hat{\pi}_1 x_1 + \hat{\pi}_2 x_2$. Explain why we can use the moment condition $\sum \hat{y}_{i2}(y_{i1} - \alpha_1 y_{i2}) = 0$ as a valid basis for consistently estimating α_1. Obtain the IV estimate of α_1.

h. Find the 2SLS estimate of α_1 by applying OLS to $y_1 = \alpha_1 \hat{y}_2 + e_1^*$. Compare your answer to that in part (g).

11.2 Consider a supply and demand model written in its most general implicit form, using capital Greek letters for the unknown parameters and E_i for the random errors,

$$\text{Demand: } \Gamma_{11}q + \Gamma_{21}p + \mathrm{B}_{11} + \mathrm{B}_{21}x + E_1 = 0$$
$$\text{Supply: } \Gamma_{12}q + \Gamma_{22}p + \mathrm{B}_{12} + \mathrm{B}_{22}x + E_2 = 0$$

a. Multiply each equation by 3. Do they remain true?

b. Multiply the demand equation by $-1/\Gamma_{11}$. Does it remain true?

c. Define $\alpha_{21} = -\Gamma_{21}/\Gamma_{11}$, $\beta_{11} = -\mathrm{B}_{11}/\Gamma_{11}$, $\beta_{21} = -\mathrm{B}_{21}/\Gamma_{11}$, $e_1 = -E_1/\Gamma_{11}$ and write the demand equation with q on the left-hand side and the remaining terms on the right-hand side. By choosing q to be on the left-hand side of the equation, we have chosen a **normalization rule**.

d. Repeat the process for the supply equation, beginning by multiplying through by $-1/\Gamma_{22}$, and obtain the normalized supply curve with

$$\alpha_{12} = -\Gamma_{12}/\Gamma_{22},\ \beta_{12} = -\mathrm{B}_{12}/\Gamma_{22},\ \beta_{22} = -\mathrm{B}_{22}/\Gamma_{22},\ \text{and}\ e_2 = -E_2/\Gamma_{22}$$

Write the normalized supply equation with p on the left-hand side and the remaining terms on the right side.

e. Mathematically, in a system of jointly determined variables, it does not matter which variable appears on the left side of each normalized equation. True or false?

11.3 Consider a supply and demand model written in its most general implicit form, using capital Greek letters for the unknown parameters and E_i for the random errors:

$$\text{Demand: } \Gamma_{11}q + \Gamma_{21}p + \mathrm{B}_{11} + \mathrm{B}_{21}x + E_1 = 0$$
$$\text{Supply: } \Gamma_{12}q + \Gamma_{22}p + \mathrm{B}_{12} + \mathrm{B}_{22}x + E_2 = 0$$

a. Find the reduced-form equation for p, $p = \pi_1 + \pi_2 x + v$. Express π_1 and π_2 in terms of parameters Γ_{ij} and B_{ij}.

b. Suppose we replace the "true" demand equation with an equation that is a mixture of the demand and supply equations, that is, multiply through the demand equation by 3 and the supply equation by 2 and then add the two equations together to obtain

$$(3\Gamma_{11} + 2\Gamma_{12})q + (3\Gamma_{21} + 2\Gamma_{22})p + (3\mathrm{B}_{11} + 2\mathrm{B}_{12}) + (3\mathrm{B}_{21} + 2\mathrm{B}_{22})x + (3E_1 + 2E_2) = 0$$

or $\Gamma'_{11}q + \Gamma'_{21}p + \mathrm{B}'_{11} + \mathrm{B}'_{21}x + E'_1 = 0$, with $'$ denoting the new parameters. Using the new demand equation, and the original supply equation, find the reduced-form equation for p, $p = \pi_1^* + \pi_2^* x + v^*$. Express π_1^* and π_2^* in terms of parameters Γ_{ij} and B_{ij}. Compare the solution to that in (a).

11.4 Consider the supply and demand model below:

$$\text{Demand: } q = -p + 3 + 2x + e_1$$
$$\text{Supply: } p = q + 1 + x + e_2$$

a. Find the reduced-form equations for p and q as a function of the exogenous variable x.
b. Now suppose that the demand equation is $q = -5p + 11 + 8x + e_1^*$. Find the reduced-form equations for p and q using this demand equation and the original supply equation.
c. Show that the new demand equation is a mixture of the original supply and demand equations. Specifically, it is three times the original demand equation plus two times the supply equation. [*Hint*: It is simpler to put the demand and supply equations into implicit form, with everything on the left side and zero on the right side, before doing the multiplying and adding.]
d. If we have N observations on p, q, and x, can we consistently estimate the demand equation by OLS? Why?
e. If we have N observations on p, q, and x, can we consistently estimate the reduced-form equations by OLS? Why?
f. Given the true reduced-form equations, can we deduce whether $q = -p + 3 + 2x + e_1$ or $q = -5p + 11 + 8x + e_1^*$ is the true demand equation?
g. Is the demand equation "identified" using the necessary condition?

11.5 Consider the supply and demand model below:

$$\text{Demand: } q = -p + 3 + 2x + e_1$$
$$\text{Supply: } p = q + 1 + e_2$$

a. Find the reduced-form equations for p and q as a function of the exogenous variable x.
b. Now suppose that the demand equation is $q = -5p + 11 + 6x + e_1^*$. Find the reduced-form equations for p and q using this demand equation and the original supply equation.
c. Show that the new demand equation is a mixture of the original supply and demand equations. Specifically, it is three times the original demand equation plus two times the supply equation. [*Hint*: It is simpler to put the demand and supply equations into implicit form, with everything on the left side and zero on the right side, before doing the multiplying and adding.]
d. If we have N observations on p, q, and x, can we consistently estimate the supply equation by OLS? Why?
e. If we have N observations on p, q, and x, can we consistently estimate the reduced-form equations by OLS? Why?
f. Given the economic supply and demand model proposed in the question, is it possible for the mixture equation $q = -5p + 11 + 6x + e_1^*$ to be a supply curve? Explain.
g. Is the demand equation "identified" using the necessary condition? Is the supply equation "identified" using the necessary condition?

11.6 Consider the supply and demand model below, where x is exogenous.

$$\text{Demand: } q = \alpha_1 p + \alpha_2 + \alpha_3 x + e_1$$
$$\text{Supply: } p = \beta_1 q + \beta_2 + e_2$$

a. Find the reduced-form equations for p and q, $q = \pi_{11} + \pi_{21}x + v_1$ and $p = \pi_{12} + \pi_{22}x + v_2$, expressing the reduced-form parameters in terms of α's and β's.
b. Suppose $\pi_{11} = 1/5$, $\pi_{21} = 3/5$, $\pi_{12} = 2/5$, and $\pi_{22} = 6/5$. Solve for as many of the α's and β's as you can.

11.7 Consider the supply and demand model below, where x and w are exogenous.

$$\text{Demand: } q = \alpha_1 p + \alpha_1 x + \alpha_2 w + e_1$$
$$\text{Supply: } p = \beta_1 q + e_2$$

a. Find the reduced-form equations for p and q, $q = \pi_{11}x + \pi_{21}w + v_1$ and $p = \pi_{12}x + \pi_{22}w + v_2$, expressing the reduced-form parameters in terms of α's and β's.
b. Suppose $\pi_{11} = 1/5$, $\pi_{21} = 1/5$, $\pi_{12} = 2/5$, and $\pi_{22} = 2/5$. Solve for as many of the α's and β's as you can.

11.8 In macroeconomics, the simple "consumption function" relates national expenditure on consumption goods, $CONSUMP_t$ = aggregate consumption, in period t to national income, $INCOME_t = GNP_t$. Specify the consumption function $CONSUMP_t = \beta_1 + \beta_2 INCOME_t + e_t$. Suppose that INV_t is aggregate investment. In the simplest model, the income identity is $INCOME_t = CONSUMP_t + INV_t$.

a. Substitute the income identity into the consumption function and solve for consumption in terms of investment.

b. Find the covariance between $INCOME_t$ and the random error e_t.

c. Find the covariance between INV_t and $INCOME_t$.

d. Suppose INV_t is uncorrelated with the random error e_t. Does it satisfy the conditions for an IV?

11.9 Consider the simultaneous equations model, where x is exogenous.

$$y_{i1} = \alpha_1 y_{i2} + \alpha_2 x_{i1} + e_{i1}$$
$$y_{i2} = \alpha_2 y_{i1} + \beta_1 x_{i1} + e_{i2}$$

Assume that $E(e_{i1}|\mathbf{x}_1) = E(e_{i2}|\mathbf{x}_1) = 0$, $\text{var}(e_{i1}|\mathbf{x}_1) = \sigma_1^2$, $\text{var}(e_{i2}|\mathbf{x}_1) = \sigma_2^2$, and $\text{cov}(e_{i1}, e_{i2}|\mathbf{x}_1) = \sigma_{12}$.

a. Substitute the second equation into the first and find the reduced-form equation for y_{i1}.

b. Multiply the reduced-form equation for y_{i1} from part (a) by e_{i2} and find $\text{cov}(y_{i1}, e_{i2}|\mathbf{x}_1) = E(y_{i1}e_{i2}|\mathbf{x}_1)$.

c. Show that $\text{cov}(y_{i1}, e_{i2}|\mathbf{x}_1) = 0$ if $\alpha_1 = 0$ and $\sigma_{12} = 0$. Such a system is said to be **recursive**.

d. Is the OLS estimator of the first equation consistent under the conditions in (c)? Explain.

e. Is the OLS estimator of the second equation consistent under the conditions in (c)? Explain.

11.10 Reconsider the Truffle supply and demand model in Example 11.1. Modify the demand equation as $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 PS_i + e_i^d$, keeping the supply equation unchanged. The estimates are given in Table 11.6.

TABLE 11.6 Estimates for Exercise 11.10

	(1)	(2)	(3)	(4)
C	5.6169	−40.5043	0.4460	19.9625
	(3.6256)	(10.0873)	(4.1596)	(1.2371)
PF	−0.2762	2.1635		−1.0425
	(0.1097)	(0.3053)		(0.0907)
PS	0.8685	2.4522	1.1815	
	(0.1434)	(0.3991)	(0.2765)	
P			−0.1277	0.3542
			(0.0671)	(0.0288)

Standard errors in parentheses.

a. Are the demand and supply equations identified using the necessary condition in Section 11.4? Explain.

b. Column (1) contains the OLS estimates of the reduced-form equation for Q, and column (2) contains the OLS estimates of the reduced-form equation for P. Compute the first-stage F-test used to decide upon instrument strength for each equation. Is the F-value greater than the rule of thumb threshold, $F > 10$? [*Hint*: Recall the relationship between t- and F-tests.]

c. Using the estimates accurately sketch the supply and demand equations, with Q on the vertical axis and P on the horizontal axis. For these sketches set the values of the exogenous variables, PS and PF, to be $PF^* = 23$ and $PS^* = 22$.

d. What are the equilibrium values of P and Q from (c)?

e. On the graph from part (c) show the consequences of increasing the price of the factor of production (the truffle-seeking pig's rental rate) from $PF^* = 23$ to $PF^* = 30$, holding the value of PS constant.

f. Calculate the change in equilibrium price P and quantity Q in (d). What is the percentage change in equilibrium quantity divided by the percentage change in PF?

g. Calculate a 95% interval estimate for the elasticity of Q with respect to PF using the reduced-form equation estimates, at $PF^* = 23$ and $PS^* = 22$. Is the elasticity in (f) within the 95% interval estimate?

11.11 Reconsider the Truffle supply and demand model in Example 11.1. Suppose we modify the supply equation to be $Q_i = \beta_1 + \beta_2 P_i + e_i^s$, keeping the demand equation unchanged.

a. Are the supply and demand equations identified using the necessary condition in Section 11.4? Explain.

b. The estimated first-stage, reduced form, equation becomes

$$\underset{(t)}{\hat{P}_i = -13.50 + 1.47PS_i + 12.41DI_i} \qquad F = 54.21$$
$$(3.23) \quad (6.95)$$

Do you judge the omitted exogenous variables (instruments) strong enough to estimate the identified equation(s)? Explain.

c. The estimated supply equation using 2SLS is

$$\underset{(\text{se})}{\hat{Q}_i = 8.6455 + 0.1564P_i}$$
$$(2.89) \quad (0.045)$$

Verify that the point of the means (see Table 11.1) falls on the estimated supply curve.

d. Calculate the price elasticity of supply at the means and compare it to the elasticity computed from the 2SLS estimates in Table 11.3b.

e. Comparing the results in parts (b) and (c) to those in Example 11.1, do you think we should include PF in the supply equation? Explain.

11.12 Suppose you want to estimate a wage equation for married women of the form

$$\ln(WAGE) = \beta_1 + \beta_2 HOURS + \beta_3 EDUC + \beta_4 EXPER + \beta_5 EXPER^2 + e_1$$

where $WAGE$ is the hourly wage, $HOURS$ is number of hours worked per week, $EDUC$ is years of education, and $EXPER$ is years of experience. Your classmate observes that higher wages can bring forth increased work effort, and that married women with small children may reduce their hours of work to take care of them. It may also be true that a husband's wage rate has an effect on a wife's hours of work supplied, so that there may be an auxiliary relationship such as

$$HOURS = \alpha_1 + \alpha_2 \ln(WAGE) + \alpha_3 KIDS + \alpha_4 \ln(HWAGE) + e_2$$

where $KIDS$ is the number of children under the age of six in the woman's household and $HWAGE$ is her husband's wage rate.

a. Can the wage equation be estimated satisfactorily using the OLS estimator? If not, why not?

b. Is the wage equation "identified"? What does the term *identification* mean in this context?

c. If you seek an alternative to least squares estimation for the wage equation, suggest an estimation procedure and how (step by step, and NOT a computer command) it is carried out.

d. Other than the identification condition in part (b), are there any other conditions that must be met so that we can confidently use the estimation procedure in part (c)? What are those conditions?

11.13 In the post-World War II period, monetary policy effects and the supply and demand for money were important topics. Consider the following model, where M is the money stock, R is short-term rate of interest, GNP is national income, and R_d is Federal Reserve's discount rate, which it charges commercial banks. The endogenous variables are the money supply, M, and the short-term rate of interest, R. The exogenous variables are GNP and the Federal Reserve's discount rate, R_d. The lagged money stock M_{t-1} is a **predetermined** variable. It is treated as an exogenous variable and uncorrelated with the current period error. Using quarterly data from the post-war period, the 2SLS estimated money demand, omitting seasonal and other dummy variables, is

$$\underset{(\text{se})}{\hat{M}_t = 23.06 + 0.0618GNP_t - 0.0025(R \times GNP_t) + 0.686M_{t-1} + \cdots}$$
$$(0.0126) \quad (0.0007) \quad (0.0728)$$

The supply equation is taken to be proportional to the difference between the short-term interest rate R and the discount rate, R_d, with the factor of proportionality being the maximum potential money

stock, M^*, which is a known constant. The estimated supply equation, omitting seasonal and other dummy variables, is

$$\hat{M}_t = 0.8522 + 0.0751\left[M_t^*\left(R_t - R_{d,t}\right)\right] + \cdots$$
$$\text{(se)} \qquad\qquad (0.0159)$$

a. If the preceding period's money supply increases by one unit, what happens to the money demand function? In a graph like Figure 11.4, with M on the vertical axis and R on the horizontal axis, does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta M_{t-1} > 0$? If so, which direction?

b. If GNP increases by one unit, what happens to the money demand function? In a graph like Figure 11.4, does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta GNP_t > 0$? If so, which direction?

c. If the discount rate, R_d, is increased does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta R_{d,t} > 0$? If so, which direction?

d. Explain how your answers to (a), (b), and (c) imply that both the supply and demand for money functions are identified.

11.14 Australian wine is popular in Australia and worldwide. Using annual data on wine grape transactions Q(10,000 tonne units) and price P($AU100 per tonne) of wine produced in warm inland Australia, an estimated demand equation is

$$\hat{Q}_t = -0.278P_t + 2.884INCOME_t - 3.131XRATE_t - 2.766STOCKS_{t-1} + \cdots$$
$$(t) \quad (-2.85) \qquad (6.34) \qquad\quad (-3.04) \qquad\quad (-2.24)$$

INCOME (US$1,000,000) is weighted household consumption expenditure, *XRATE* is the exchange rate per $AU, and *STOCKS* (1000 million litres) are from the previous year. An estimated supply equation is

$$\hat{Q}_t = 0.824P_t + 0.682Q_{t-4} + 0.598TIME_t - 1.688TEMP_t + 1.793NON_{t-4} - 1.570PREM_{t-4} + \cdots$$
$$(t) \quad (4.82) \qquad (3.68) \qquad (5.87) \qquad (-1.19) \qquad\quad (4.21) \qquad\quad (-2.42)$$

TEMP is mean January temperature (mid-summer "Down Under"), NON_{t-4} is the price of the regional wine grape relative to other non-premium grapes, lagged four years, and $PREM_{t-4}$ is the price of the regional wine grapes relative to other premium wine grapes, lagged four years. Production at time $t-4$ is on the right-hand side reflecting the four years required between planting grape vines and producing wine. This is a **partial adjustment** model as discussed in Exercise 9.30. In both equations, we have omitted the intercept and indicator variables for specific regions.

a. Which variables in the model cause the demand equation to shift relative to the supply equation?

b. Which variables in the model cause the supply equation to shift relative to the demand equation?

c. Discuss the signs of the estimated coefficients in the demand equation.

d. Sample means of Q, P, and *INCOME* are $\overline{Q} = 4.98$, $\overline{P} = 6.06$, and $\overline{INCOME} = 1.66$. Calculate the price and income elasticity of demand at the means.

e. Discuss the signs of the estimated coefficients in the supply equation.

f. Calculate the elasticity of equilibrium supply with respect to price at the means.

11.15 Consider the supply and demand for labor, and in particular that for married women. Wages and hours worked are jointly determined by supply and demand. Let the supply equation be

$$HOURS = \beta_1 + \beta_2 \ln(WAGE) + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 KIDS618 + \beta_7 NWIFEINC + e_e$$

KIDSL6 are the number of children less than 6 years old, *KIDS618* are the number of children who are 6 to 18 years old, *NWIFEINC* is household income other than the wife's earnings. Let the demand equation be

$$\ln(WAGE) = \alpha_1 + \alpha_2 HOURS + \alpha_3 EDUC + \alpha_4 EXPER + \alpha_5 EXPER^2 + e_d$$

a. Imagine a supply and demand graph, like Figure 11.4, with *HOURS* on the vertical axis and ln(*WAGE*) on the horizontal axis. Describe the anticipated effects on the graph of increases in the number of small children on the woman's supply and demand curves. What is the anticipated effect on equilibrium wage and hours worked?

b. Describe the anticipated effects on the graph of increases in experience on the woman's supply and demand curves. What is the anticipated effect on equilibrium wage and hours worked?

c. Does the necessary condition for identification appear to hold for the supply equation? What are the IVs used in 2SLS? Write out the econometric form of the reduced-form equation for ln(*WAGE*), letting the coefficients be denoted as π_1, π_2, etc. What hypotheses would you test to evaluate the strength of IVs used in 2SLS estimation of the supply equation?

d. Does the necessary condition for identification appear to hold for the demand equation? What are the IVs used in 2SLS? Write out the econometric form of the reduced-form equation for *HOURS*, letting the coefficients be denoted as γ_1, γ_2, etc. What hypotheses would you test to evaluate the strength of IVs used in 2SLS estimation of the demand equation?

11.16 Consider the following supply and demand model

$$\text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + e_{di}, \quad \text{Supply: } Q_i = \beta_1 + \beta_2 P_i + \beta_3 W_i + e_{si}$$

where Q is the quantity, P is the price, and W is the wage rate, which is assumed exogenous. Data on these variables are in Table 11.7.

TABLE 11.7 Data for Exercise 11.16

Q	P	W
4	2	2
6	4	3
9	3	1
3	5	1
8	8	3

a. Derive the algebraic form of the reduced-form equations, $Q = \theta_1 + \theta_2 W + v_2$ and $P = \pi_1 + \pi_2 W + v_1$, expressing the reduced-form parameters in terms of the structural parameters.

b. Which structural parameters can you solve for from the results in part (a)? Which equation is "identified"?

c. The estimated reduced-form equations are $\hat{Q} = 5 + 0.5W$ and $\hat{P} = 2.4 + 1W$. Solve for the identified structural parameters. This is the method of **indirect least squares**.

d. Obtain the fitted values from the reduced-form equation for P, and apply 2SLS to obtain estimates of the demand equation.

11.17 Example 11.3 introduces Klein's Model I.

a. Do we have an adequate number of IVs to estimate each equation? Check the necessary condition for the identification of each equation. The necessary condition for identification is that in a system of M equations at least $M - 1$ variables must be omitted from each equation.

b. An equivalent identification condition is that the number of excluded exogenous variables from the equation must be at least as large as the number of included right-hand side endogenous variables. Check that this condition is satisfied for each equation.

c. Write down in econometric notation the first-stage equation, the reduced form, for W_{1t}, wages of workers earned in the private sector. Call the parameters π_1, π_2,…

d. Describe the two regression steps of 2SLS estimation of the consumption function. This is not a question about a computer software command.

e. Does following the steps in part (d) produce regression results that are identical to the 2SLS estimates provided by software specifically designed for 2SLS estimation? In particular, will the t-values be the same?

11.6.2 Computer Exercises

11.18 Example 11.3 introduces Klein's Model I. Here we examine a simplified model that excludes the government sector and allows further practice with simultaneous equations models. Suppose the model is

reduced to the following two equations, for two endogenous variables, consumption, *CN*, and investment, *I*. The two estimable equations are the consumption and investment functions:

$$CN_t = \alpha_1 + \alpha_2 I_t + \alpha_3 TIME_t + e_{1t}$$

$$I_t = \beta_1 + \beta_2 CN_t + \beta_3 K_{t-1} + e_{2t}$$

a. Check the identification of the consumption and investment functions.

b. Solve for the reduced-form equation for *CN*. Call the parameters π_1, π_2, π_3 and express them in terms of the structural parameters, similar to equations (11.4) and (11.5).

c. Using the data file *klein*, estimate each of the structural equations by OLS. Comment on the signs and significance of the coefficients.

d. Estimate each of the structural equations by 2SLS. Comment on the signs and significance of the coefficients.

e. Estimate the first-stage, reduced form, equation. In the reduced-form equation for consumption is K_{t-1} statistically significant? In the reduced-form equation for investment is $TIME_t$ statistically significant? Do these results help explain the differences in the OLS and 2SLS estimates?

11.19 The labor supply of married women has been a subject of a great deal of economic research. The data file is *mroz*, and the variable definitions are in the file *mroz.def*. The data file contains information on women who have worked in the previous year and those who have not. The variable indicating whether a woman worked *LFP*, labor force participation, takes the value 1 if a woman worked and 0 if she did not.

a. Calculate the summary statistics for the variables: wife's age, the number of less than 6-year-old children, and the income from other sources than from the wife's employment, *NWIFEINC*, for the women who worked (*LFP* = 1) and those who did not (*LFP* = 0). Define *NWIFEINC* = *FAMINC* – *WAGE* × *HOURS*. Comment on any differences you observe.

b. Consider the following supply equation specification:

$$HOURS = \beta_1 + \beta_2 \ln(WAGE) + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 KIDS618 + \beta_7 NWIFEINC + e$$

What signs do you expect each of the coefficients to have, and why? What does *NWIFEINC* measure?

c. Estimate the supply equation in (b) using OLS regression on *only the women who worked* (*LFP* = 1). Did things come out as expected? If not, why not?

d. Estimate the reduced-form equation by OLS for the women who worked, using work experience, *EXPER*, as an additional exogenous variable.

$$\ln(WAGE) = \pi_1 + \pi_2 EDUC + \pi_3 AGE + \pi_4 KIDSL6 + \pi_5 KIDS618 + \pi_6 NWIFEINC + \pi_7 EXPER + v$$

Based on the estimated reduced form, what is the effect upon wage of an additional year of education?

e. Check the identification of the supply equation, considering the availability of instrument *EXPER*.

f. Estimate the supply equation by two-stage least squares, using software designed for this purpose. Discuss the signs and significance of the estimated coefficients.

11.20 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We consider the demand equation in this exercise and the supply equation in Exercise 11.21.

a. Consider the demand equation:

$$\ln(Q_t) = \alpha_1 + \alpha_2 \ln(P_t) + \alpha_3 \ln(Y_t) + \alpha_4 \ln(PB_t) + \alpha_5 POPGRO_t + e_t^d$$

where Q = per capita consumption of chicken, in pounds; Y = real per capita income; P = real price of chicken; PB = real price of beef; and $POPGRO$ = rate of population growth. What are the endogenous variables? What are the exogenous variables?

b. Using data from 1960 to 1999, estimate the demand equation by OLS. Comment on the signs and significance of the estimates.

c. Test the OLS residuals from part (b) for serial correlation by constructing a correlogram and carrying out the $T \times R^2$ test. What do you conclude about the presence of serial correlation?

d. Estimate the demand equation by 2SLS using as instruments $\ln(PF_t)$, $TIME_t = YEAR_t - 1949$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Compare and contrast these estimates to the OLS estimates in part (a).

e. Estimate the reduced-form, first-stage, equation and test the joint significance of $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong?

f. Test the reduced-form equation for serial correlation using the $T \times R^2$ test.

g. Estimate the reduced-form, first-stage, equation using HAC standard errors and test the joint significance of $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$.

h. Obtain the 2SLS residuals from part (d). Construct a correlogram. Is there evidence of serial correlation? Obtain 2SLS estimates with HAC standard errors and compare the results to those in (d).

i. Test the validity of the surplus instruments using the Sargan test, discussed in Section 10.4.3, and the 2SLS estimates in part (d).

11.21 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We considered the demand equation in Exercise 11.20. The supply equation is

$$\ln(QPROD_t) = \beta_1 + \beta_2 \ln(P_t) + \beta_3 \ln(PF_t) + \beta_4 TIME_t + \ln(QPROD_{t-1}) + e_t^s$$

where *QPROD* is the aggregate production of young chickens, *PF* is nominal price index of broiler feed, and *TIME* = time index with $1950 = 1, \ldots, 2001 = 52$. This supply equation is dynamic, with lagged production on the right-hand side. This predetermined variable is exogenous. *TIME* is included to capture technical progress in production.

a. What are the endogenous variables? What are the exogenous variables? What is the interpretation of the parameter β_2? What signs do you expect for each of the parameters?

b. Using data from 1960 to 1999, estimate the supply equation by OLS. Comment on the signs and significance of the estimates. Test the residuals for serial correlation. Is serial correlation present?

c. Estimate the reduced-form, first-stage, regression by OLS using the IVs $\ln(Y_t)$, $\ln(PB_t)$, *POPGRO*, and $\ln(EXPTS_{t-1})$. Test the joint significance of these variables. Can we conclude that we have at least one strong instrument?

d. Estimate the supply equation by 2SLS using the instruments listed in part (c). Compare and contrast these results to those in part (b).

e. Test the validity of the surplus instruments using the Sargan test, discussed in Section 10.4.3.

11.22 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We considered the demand equation in Exercise 11.20. It is

$$\ln(Q_t) = \alpha_1 + \alpha_2 \ln(P_t) + \alpha_3 \ln(Y_t) + \alpha_4 \ln(PB_t) + \alpha_5 POPGRO_t + e_t^d$$

where Q is the per capita consumption of chicken, in pounds; Y is real per capita income; P is real price of chicken; *PB* is real price of beef, and *POPGRO* is rate of population growth. What are the endogenous variables? What are the exogenous variables? The demand equation suffers from serial correlation. In the AR(1) model $e_t^d = \rho e_{t-1}^d + v_t^d$ the value of ρ is large. Epple and McCallum estimate the model in "first difference" form:

$$\begin{aligned} \ln(Q_t) &= \alpha_1 + \alpha_2 \ln(Y_t) + \alpha_3 \ln(P_t) + \alpha_4 \ln(PB_t) + e_t^d \\ -\left[\ln(Q_t) = \alpha_1 + \alpha_2 \ln(Y_t) + \alpha_3 \ln(P_t) + \alpha_4 \ln(PB_t) + e_t^d\right] & \\ \hline \Delta \ln(Q_t) = \alpha_2 \Delta \ln(Y_t) + \alpha_3 \Delta \ln(P_t) + \alpha_4 \Delta \ln(PB_t) + v_t^d & \end{aligned}$$

a. Regarding this specification (i) what changes do you notice after this transformation? (ii) Are the parameters of interest affected? (iii) If $\rho = 1$, have we solved the serial correlation problem? (iv) What is the interpretation of the "Δ" variables like $\Delta \ln(Q_t)$? [*Hint*: See Appendix A.4.6.] (v) What is the interpretation of the parameter α_2? (vi) What signs do you expect for each of the coefficients? Explain.

b. Using data from 1960 to 1999, estimate the reduced-form, first-stage, equation for $\Delta\ln(P_t)$ using instruments $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong?

c. Estimate the first-stage equation for $\Delta\ln(P_t)$ using instruments $\Delta\ln(PF_t)$, $\Delta\ln(QPROD_{t-1})$, and $\Delta\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong? On logical grounds, why might we prefer these instruments to those in (b)?

d. Estimate the first-stage equation for $\Delta\ln(P_t)$ using instrument $\Delta\ln(PF_t)$. Can we conclude that the one instrument is strong?

e. Obtain the 2SLS estimates of the first-differenced demand equation using $\Delta\ln(PF_t)$ as the instrument. In this estimation omit the constant term.

f. Obtain the 2SLS estimates of the first-differenced demand equation using $\Delta\ln(PF_t)$ as the instrument including a constant term.

g. Compare the estimates of the key demand parameters in parts (e) and (f). Are the signs consistent with expectations? What are the interpretations of the estimated coefficients? Should an intercept be included in the differenced demand equation? Explain.

h. Construct a correlogram for the 2SLS residuals in part (e). Is there any evidence of serial correlation?

11.23 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*.

a. Obtain OLS estimates of the supply equation. Comment on the coefficient signs and significance. Do you anticipate the OLS estimator to have a positive bias or a negative bias or no bias? Explain.

b. It is possible that bad weather on shore reduces attendance at restaurants, which in turn may reduce the demand for fish at the market. Add the variables *RAINY* and *COLD* to the demand equation in (11.13). Derive the algebraic reduced form for ln(*PRICE*) for this new specification.

c. Estimate the reduced-form equation in part (b). Test the joint significance of *RAINY* and *COLD*. Are these variables jointly significant at the 5% level?

d. Using the estimates from part (c), test the joint significance of *MON*, *TUE*, *WED*, *THU*, *RAINY*, and *COLD*. Are these variables jointly significant at the $\alpha = 0.05$ level?

e. Estimate the supply equation by 2SLS using instruments *MON*, *TUE*, *WED*, *THU*, *RAINY*, and *COLD*. Compare these estimates to the OLS estimates in part (a). Given the results in part (d), can we conclude that the supply equation is identified?

11.24 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*.

a. Add the variable *MIXED*, which indicates poor but not *STORMY* weather conditions, to the supply equation in equation (11.14). Estimate the new reduced-form equation for ln(*PRICE*), adding the variable *MIXED* to equation (11.16). Is it statistically significant at the 5% level? Test the joint significance of *STORMY* and *MIXED*. Is the resulting *F*-value greater than 10?

b. Estimate the demand equation using *STORMY* and *MIXED* as IVs. Compare the coefficient estimates to those in Table 11.5.

c. Test the validity of the surplus instrument using the Sargan test, discussed in Section 10.3.4.

d. In the reduced-form equation in part (a), test the joint significance of the indicator variables *MON*, *TUE*, *WED*, and *THU* at the 5% level. What do you conclude? Are we now able to estimate the supply equation by 2SLS with confidence in our procedure?

11.25 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*. In this exercise, we explore the behavior of the market on days in which changes in fish inventories are large relative to those days on which inventory changes are small. Graddy and Kennedy (2006) anticipate that prices and quantities will demonstrate simultaneity on days with large changes in inventories, as these are days when sellers are demonstrating their responsiveness to prices. On days when inventory changes are small, the anticipation is that feedback between prices and quantities is broken, and simultaneity is no longer an issue.

a. Use the subset of data for days in which inventory change is large, as indicated by the variable $CHANGE = 1$. Estimate the reduced-form equation (11.16) and test the significance of *STORMY*. Discuss the importance of this test for the purpose of estimating the demand equation by two-stage least squares.

b. Obtain the OLS residuals $\hat{v}_{t2}$ from the reduced-form equation estimated in (a). Carry out a Hausman test, as discussed in Section 10.4.1, for the endogeneity of ln(*PRICE*) by adding $\hat{v}_{t2}$

as an extra variable to the demand equation in (11.13), estimating the resulting model by OLS, and testing the significance of $\hat{v}_{t2}$ using a standard t-test. If $\hat{v}_{t2}$ is a significant variable in this augmented regression then we may conclude that ln(*PRICE*) is endogenous. Based on this test, what do you conclude?

c. Estimate the demand equation using two-stage least squares and OLS using the data when $CHANGE = 1$, and discuss these estimates. Compare them to the estimates in Table 11.5.

d. Estimate the reduced-form equation (11.16) for the data when $CHANGE = 0$. Compare these reduced-form estimates to those in (a) and those in Table 11.4b.

e. Obtain the OLS residuals $\hat{v}_{t2}$ from the reduced-form equation estimated in (d). Carry out a Hausman test for the endogeneity of ln(*PRICE*), as described in part (b). Based on this test, what do you conclude?

f. Obtain the two-stage least squares and the OLS estimates for the demand equation for the data when $CHANGE = 0$. Compare these estimates to each other and to the estimates in (c). Discuss the relationships between them.

11.26 Use your computer software for two-stage least squares or IVs estimation, and the 30 observations in the data file *truffles* to obtain 2SLS estimates of the system in equations (11.4) and (11.5). Compare your results to those in Tables 11.3a and 11.3b.

a. Using the 2SLS estimated equations, compute the price elasticity of supply and demand "at the means." Comment on the signs and magnitudes of these elasticities.

b. Using the 2SLS estimates for the demand equation, obtain the squared 2SLS residuals, $\hat{e}_d^2$. Carry out the Breusch–Pagan NR^2 test for heteroskedasticity using just the exogenous variables in the variance function. Is there any evidence of heteroskedasticity?

c. Using the 2SLS estimates for the supply equation, obtain the squared 2SLS residuals, $\hat{e}_s^2$. Carry out the Breusch–Pagan NR^2 test for heteroskedasticity using just the exogenous variables in the variance function. Is there any evidence of heteroskedasticity?

d. Plot the squared supply equation residuals $\hat{e}_s^2$ versus each of the three exogenous variables. Discuss the visual evidence of heteroskedasticity.

e. Obtain 2SLS estimates of the supply equation using robust standard errors. How do the t-statistic values compare to those in Table 11.3b? Do you think it is a good idea to use robust standard errors for this equation? Explain.

11.27 Estimate equations (11.4) and (11.5) by OLS, ignoring the fact that they form a simultaneous system. Use the data file *truffles*. Compare your results to those in Table 11.3. Do the signs of the least squares estimates agree with economic reasoning?

11.28 Supply and demand curves as traditionally drawn in economics principles classes have price (P) on the vertical axis and quantity (Q) on the horizontal axis.

a. Rewrite the truffle demand and supply equations in (11.11) and (11.12) with price P on the left-hand side. What are the anticipated signs of the parameters in this rewritten system of equations?

b. Using the data in the file *truffles*, estimate the supply and demand equations that you have formulated in (a) using two-stage least squares. Are the signs correct? Are the estimated coefficients significantly different from zero?

c. Estimate the price elasticity of demand "at the means" using the results from (b).

d. Accurately sketch the supply and demand equations, with P on the vertical axis and Q on the horizontal axis, using the estimates from part (b). For these sketches set the values of the exogenous variables DI, PS, and PF to be $DI^* = 3.5$, $PF^* = 23$, and $PS^* = 22$.

e. What are the equilibrium values of P and Q obtained in part (d)? Calculate the predicted equilibrium values of P and Q using the estimated reduced-form equations from Table 11.2, using the same values of the exogenous variables. How well do they agree?

f. Estimate the supply and demand equations that you have formulated in (a) using OLS. Are the signs correct? Are the estimated coefficients significantly different from zero? Compare the results to those in part (b).

11.29 Example 11.3 introduces Klein's Model I. Use the data file *klein* to answer the following questions.

a. Estimate the consumption function in equation (11.17) by OLS. Comment on the signs and significance of the coefficients.

b. Estimate the reduced-form equation for wages of workers in the private sector, W_{1t}, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all

the variables except wages of workers in the public sector, W_{2t}, and lagged profits, P_{t-1}. Save the residuals, $\hat{v}_{1t}$.

c. Estimate the reduced-form equation for profits, P_t, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all the variables except wages of workers in the public sector, W_{2t}, and lagged profits, P_{t-1}. Save the residuals, $\hat{v}_{2t}$.

d. The Hausman test for the presence of endogenous explanatory variables is discussed in Section 10.4.1. It is implemented by adding the reduced-form residuals to the structural equation and testing their significance, that is, using OLS, estimate the model

$$CN_t = \alpha_1 + \alpha_2(W_{1t} + W_{2t}) + \alpha_3 P_t + \alpha_4 P_{t-1} + \delta_1 \hat{v}_{1t} + \delta_2 \hat{v}_{2t} + e_{1t}$$

Use an F-test for the null hypothesis $H_0: \delta_1 = 0, \delta_2 = 0$ at the 5% level of significance. By rejecting the null hypothesis, we conclude that either W_{1t} or P_t is endogenous, or both are endogenous. What do we conclude from the test? In the context of this simultaneous equations model what result should we find?

e. Obtain the 2SLS estimates of the consumption equation using all eight exogenous and predetermined variables as IVs. Compare the estimates to the OLS estimates in part (a). Do you find any important differences?

f. Let the 2SLS residuals from part (e) be $\hat{e}_{1t}$. Regress these residuals on all the exogenous and predetermined variables. If these instruments are valid, then the R^2 from this regression should be low, and none of the variables are statistically significant. The Sargan test for instrument validity is discussed in Section 10.4.3. The test statistic TR^2 has a chi-square distribution with degrees of freedom equal to the number of "surplus" IVs if the surplus instruments are valid. The consumption equation includes three exogenous and/or predetermined variables of the total of eight possible. There are $L = 5$ external instruments and $B = 2$ right-hand side endogenous variables. Compare the value of the test statistic to the 95th percentile value from the $\chi^2_{(3)}$ distribution. What do we conclude about the validity of the surplus instruments in this case?

11.30 Example 11.3 introduces Klein's Model I. Use the data file *klein* to answer the following questions.

a. Estimate the investment function in equation (11.18) by OLS. Comment on the signs and significance of the coefficients.

b. Estimate the reduced-form equation for profits, P_t, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all the variables except lagged profits, P_{t-1}, and lagged capital stock, K_{t-1}. Save the residuals, $\hat{v}_t$ and compute the fitted values, $\hat{P}_t$.

c. The Hausman test for the presence of endogenous explanatory variables is discussed in Section 10.4.1. It is implemented by adding the reduced-form residuals to the structural equation and testing their significance, that is, using OLS estimate the model

$$I_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + \delta \hat{v}_t + e_{2t}$$

Use a t-test for the null hypothesis $H_0: \delta = 0$ versus $H_1: \delta \neq 0$ at the 5% level of significance. By rejecting the null hypothesis, we conclude that P_t is endogenous. What do we conclude from the test? In the context of this simultaneous equations model what result should we find?

d. Obtain the 2SLS estimates of the investment equation using all eight exogenous and predetermined variables as IVs and software designed for 2SLS. Compare the estimates to the OLS estimates in part (a). Do you find any important differences?

e. Estimate the second-stage model $I_t = \beta_1 + \beta_2 \hat{P}_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + e_{2t}$ by OLS. Compare the estimates and standard errors from this estimation to those in part (d). What differences are there?

f. Let the 2SLS residuals from part (e) be $\hat{e}_{2t}$. Regress these residuals on all the exogenous and predetermined variables. If these instruments are valid, then the R^2 from this regression should be low, and none of the variables are statistically significant. The Sargan test for instrument validity is discussed in Section 10.4.3. The test statistic TR^2 has a chi-square distribution with degrees of freedom equal to the number of "surplus" IVs if the surplus instruments are valid. The investment equation includes three exogenous and/or predetermined variables out of the total of eight possible. There are $L = 5$ external instruments and $B = 1$ right-hand side endogenous variables. Compare the value of the test statistic to the 95th percentile value from the $\chi^2_{(4)}$ distribution. What do we conclude about the validity of the surplus instruments in this case?

Appendix 11A 2SLS Alternatives

There has always been great interest in alternatives to the standard IV/2SLS estimator. The search for better alternatives was energized by the discovery of the problems weak instruments pose for the usual IV/2SLS estimator. In this appendix, we examine a few alternative estimators for a single equation with endogenous regressors. The equation might be part of a simultaneous equations system, or a standalone equation with an endogenous regressor, as we studied in Chapter 10. The limited information maximum likelihood (LIML) estimator was first derived by Anderson and Rubin in 1949.[1] It has played a "back seat" role relative to 2SLS over the years, but this is no longer true. There is renewed interest in LIML in the presence of weak instruments. Several modifications of LIML have been suggested by Fuller (1977) and others. These estimators are unified in a common framework, along with 2SLS, using the idea of a ***k*-class** of estimators. Later in this appendix, we provide Stock–Yogo tables of critical values for weak instruments that apply to the LIML estimator and Fuller modifications. What is illustrated by these tables is that LIML suffers less from test size aberrations than the 2SLS estimator, and that the Fuller modification suffers less from bias.

11A.1 The *k*-Class of Estimators

In a system of M simultaneous equations let the endogenous variables be $y_1, y_2, \ldots, y_M$. Let there be K exogenous variables, $x_1, x_2, \ldots, x_K$. Suppose the first structural equation within this system is

$$y_1 = \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1 \tag{11A.1}$$

If this equation is identified, then its parameters can be estimated. The variable y_2 is endogenous because it is correlated with the regression error term e_1. The endogenous variable y_2 has reduced form $y_2 = \pi_{12}x_1 + \pi_{22}x_2 + \cdots + \pi_{K2}x_K + v_2 = E(y_2|\mathbf{X}) + v_2$. The source of the endogeneity of y_2 is not the systematic portion $E(y_2|\mathbf{X})$, which is exogenous. The random component v_2 is the source of the endogeneity problem. One way to think about developing an IV for y_2 is to remove, or "purge," v_2 from it, that is, use the IV $y_2 - v_2 = E(y_2|\mathbf{X})$. This instrument has the essential properties of an instrument: It is correlated with the endogenous variable y_2 and it is uncorrelated with the structural equation error e_1. The difficulty is that $E(y_2|\mathbf{X})$ is unknown. However, the parameters of the reduced-form equation are consistently estimated by OLS, so that

$$\widehat{E(y_2|\mathbf{X})} = \hat{\pi}_{12}x_1 + \hat{\pi}_{22}x_2 + \cdots + \hat{\pi}_{K2}x_K \tag{11A.2}$$

The reduced-form residuals are

$$\hat{v}_2 = y_2 - \widehat{E(y_2|\mathbf{X})}$$

In large samples the reduced-form estimators $\hat{\pi}_{k2}$ converge in probability to their true values. This means that in large samples we can substitute for $E(y_2|\mathbf{X})$ its estimated value

$$\widehat{E(y_2|\mathbf{X})} = y_2 - \hat{v}_2 \tag{11A.3}$$

The two-stage least squares estimator is an IV estimator using $\widehat{E(y_2|\mathbf{X})}$ as an instrument. Equation (11A.3) shows that the instrument used in 2SLS can be thought of as the endogenous variable y_2 "purged" of the troublesome error term v_2.

The ***k*-class** of estimators is a unifying framework. A *k-class* estimator is an IV estimator using IV $y_2 - k\hat{v}_2$. It is called a *class* of estimators because it represents the OLS estimator if

[1] Anderson, T.W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 46–63.

$k = 0$ and the 2SLS estimator if $k = 1$. Why would we be interested in using values of k other than 1? Hopefully by adjusting this value we can improve upon the performance of the k-class estimator relative to the 2SLS estimator.

11A.2 The LIML Estimator

As noted earlier, the LIML estimator is one of the oldest estimators for an equation within a system of simultaneous equations, or any equation with an endogenous variable on the right-hand side. Rather than obtaining the LIML estimates by maximizing a likelihood function (see Appendix C.8 for an introduction to maximum likelihood estimation) we will exploit the fact that the LIML estimator is a member of the *k-class*.

The equation $y_1 = \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1$ is in **normalized form**, meaning that we have chosen one variable to appear as the dependent variable. In general the first equation can be written in **implicit form** as $\alpha_1 y_1 + \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1 = 0$. There is no rule that says y_1 has to be the dependent variable in the first equation. **Normalization** amounts to setting α_1 or α_2 to the value -1. One parameter α_i must be set to -1 so that we can identify the equation, but it does not matter which one. Let $y^* = \alpha_1 y_1 + \alpha_2 y_2$, then the unnormalized equation can be written as $y^* + \beta_1 x_1 + \beta_2 x_2 + e_1 = 0$, or

$$y^* = -\beta_1 x_1 - \beta_2 x_2 - e_1 = \theta_1 x_1 + \theta_2 x_2 + \eta \tag{11A.4}$$

In (11A.1) the exogenous variables $x_3, \ldots, x_K$ were omitted. If we had included them, (11A.4) would be

$$y^* = \theta_1 x_1 + \cdots + \theta_K x_K + \eta \tag{11A.5}$$

The **least variance ratio** estimator chooses α_1 and α_2 so that the ratio of the sum of squared residuals from (11A.4) relative to the sum of squared residuals from (11A.5) is as small as possible. Define the ratio of sum of squared residuals from the two models as

$$\ell = \frac{SSE \text{ from regression of } y^* \text{ on } x_1, x_2}{SSE \text{ from regression of } y^* \text{ on } x_1, \ldots, x_K} \geq 1 \tag{11A.6}$$

We assume that the variables $x_3, \ldots, x_K$ were omitted from (11A.1) for a reason based in economic theory. The estimates of α_1 and α_2, one of which will be set to -1, should be chosen so to make the reduced regression (11A.4) fit the data as well as possible while still imposing the condition that $x_3, \ldots, x_K$ are omitted.

The algebra required for the solution is beyond the scope of this book.[2] The interesting result is that the minimum value of ℓ in (11A.6), call it $\hat{\ell}$, results in the LIML estimator when used as k in the *k-class* estimator, that is, use $k = \hat{\ell}$ when forming the instrument $y_2 - k\hat{v}_2$, and the resulting IV estimator is the LIML estimator.

11A.2.1 Fuller-Modified LIML

A modification suggested by Wayne Fuller (1977)[3] uses the *k-class* value

$$k = \hat{\ell} - \frac{a}{N - K} \tag{11A.7}$$

where K is the total number of IVs (included and excluded exogenous variables) and N is the sample size. The value of a is a constant. Fuller says (1977, p. 951), "If one desires estimates that are nearly unbiased 'a' is set equal to 1. Presumably 'a' = 1 would be used when one is interested in testing hypotheses or setting approximate confidence intervals for the parameters." Fuller also showed that the value $a = 4$ leads to an estimator that minimizes the "mean square

[2] Advanced students should consider reading Peter Schmidt's *Econometrics*, 1976, Chapter 4, New York, NY: Marcel Dekker. Inc.

[3] Wayne Fuller, "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953.

error" of estimation. If we are estimating some parameter δ using an estimator $\hat{\delta}$, then the mean square error of estimation is

$$MSE\left(\hat{\delta}\right) = E\left(\hat{\delta} - \delta\right)^2 = \text{var}\left(\hat{\delta}\right) + \left[E\left(\hat{\delta}\right) - \delta\right]^2 = \text{var}\left(\hat{\delta}\right) + \left[bias\left(\hat{\delta}\right)\right]^2$$

Estimator *MSE* combines both variance and bias into a single measure.

11A.2.2 Advantages of LIML

A great deal of research has been devoted to the performance of the LIML estimator relative to the 2SLS estimator when instruments are weak and/or there are a large number of instruments. Stock and Yogo (2005, p. 106) say, "Our findings support the view that LIML is far superior to *(2)SLS* when the researcher has weak instruments …" when using interval estimates' coverage rate as the criterion. Also "… the Fuller-*k* estimator is more robust to weak instruments than *(2)SLS* when viewed from the perspective of bias." Some other findings are discussed by Mariano (2001)[4]:

- For the 2SLS estimator the amount of bias is an increasing function of the degree of over identification. The distributions of the 2SLS and least squares estimators tend to become similar when overidentification is large. LIML has the advantage over 2SLS when there are a large number of instruments.
- The LIML estimator converges to normality faster than the 2SLS estimator and is generally more symmetric.

11A.2.3 Stock–Yogo Weak IV Tests for LIML

Tables 11A.1 and 11A.2 contain Stock–Yogo critical values for testing weak instruments. These tests are discussed in Chapter 10, Appendix A. Table 11A.1 contains the critical values using the criterion of maximum LIML test size for a 5% test. Note that for $L > 1$, LIML critical values are lower than the 2SLS critical values in Table 10A.1. This means that the Cragg–Donald F-test statistic does not have to be as large for us to reject the null hypothesis that the instruments are weak when using LIML instead of 2SLS. Table 11A.2 contains the critical values for the test of weak instruments using the relative bias criterion for the Fuller modification of LIML, using $a = 1$. There is no similar table for LIML, because the LIML estimator does not have a finite expected value, and thus the concept of bias breaks down.

TABLE 11A.1 **Critical Values for the Weak Instrument Test Based on LIML Test Size (5% Level of Significance)[5]**

	B = 1				*B = 2*			
	Maximum Test Size				Maximum Test Size			
L	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	8.68	5.33	4.42	3.92	7.03	4.58	3.95	3.63
3	6.46	4.36	3.69	3.32	5.44	3.81	3.32	3.09
4	5.44	3.87	3.30	2.98	4.72	3.39	2.99	2.79

[4]Mariano, R. S. (2001), "Simultaneous equation model estimators," in *The Companion to Theoretical Econometrics*, Badi Baltagi ed., Oxford: Blackwell Publishing, pp. 139–142.

[5]These values are from Table 5.4, page 103, in Stock and Yogo (2005), *op. cit*. The authors thank James Stock and Motohiro Yogo for permission to use these results. Their tables are more extensive than the ones we provide. The significance level of the test for weak instruments is 5%.

TABLE 11A.2 **Critical Values for the Weak Instrument Test Based on Fuller-*k* Relative Bias (5% Level of Significance)[6]**

	$B = 1$				$B = 2$			
	Maximum Relative Bias				Maximum Relative Bias			
L	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
1	24.09	19.36	15.64	12.71				
2	13.46	10.89	9.00	7.49	15.50	12.55	9.72	8.03
3	9.61	7.90	6.61	5.60	10.83	8.96	7.18	6.15
4	7.63	6.37	5.38	4.63	8.53	7.15	5.85	5.10

EXAMPLE 11.4 | Testing for Weak Instruments Using LIML

This illustration was introduced in Example 10.8. With the Mroz data we estimate the *HOURS* supply equation

$$HOURS = \beta_1 + \beta_2 MTR + \beta_3 EDUC + \beta_4 KIDSL6 + \beta_5 NWIFEINC + e \quad (11A.8)$$

The reduced-form estimates are in Table 10A.3. The LIML estimates are given in Table 11A.3. The models we consider are as follows:

Model 1: endogenous: *MTR*; IV: *EXPER*

Model 2: endogenous: *MTR*; IV: *EXPER*, $EXPER^2$, *LARGECITY*

Model 3: endogenous: *MTR*, *EDUC*; IV: *MOTHEREDUC*, *FATHEREDUC*

Model 4: endogenous: *MTR*, *EDUC*; IV: *MOTHEREDUC*, *FATHEREDUC*, *EXPER*

TABLE 11A.3 **LIML Estimations**

MODEL	(1)	(2)	(3)	(4)
C	17423.7211	16191.3338	−24491.5972	18587.9064
	(5.56)	(5.40)	(−0.31)	(5.05)
MTR	−18456.5896	−17023.8164	29709.4652	−19196.5172
	(−5.08)	(−4.90)	(0.33)	(−4.79)
EDUC	−145.2928	−134.5504	258.5590	−197.2591
	(−4.40)	(−4.26)	(0.31)	(−3.05)
KIDSL6	151.0229	113.5034	−1144.4778	207.5531
	(1.07)	(0.84)	(−0.46)	(1.27)
NWIFEINC	−103.8983	−96.2895	149.2325	−104.9415
	(−5.27)	(−5.11)	(0.32)	(−5.07)
N	428	428	428	428
$\hat{\ell}$	1.0000	1.0195	1.0000	1.0029
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60
NUMBER IV *L*	1	3	2	3
NUMBER ENDOG *B*	1	1	2	2

t-statistics in parentheses.

[6]These values are from Table 5.3, page 102, in James H. Stock and Motohiro Yogo (2005), *op. cit.*

First, for the just identified equations for which the number of instruments equals the number of endogenous variables in Models (1) and (3), the LIML estimates are identical to the 2SLS estimators. This identity is always true for just-identified equations. For the overidentified Models (2) and (4), the estimated values $\hat{\ell}$ are close to 1, so that the estimates are not too far from the 2SLS estimates.

The estimates are not the important aspect of this illustration. The Cragg–Donald F-test statistic is the same for all the estimators. For convenience its values for each equation are given at the bottom of Table 11A.3. In Model (2), we have $B = 1$ endogenous variable and $L = 3$ instruments. Using the LIML maximum size of 10% as our criterion, the Stock–Yogo critical value is 6.46. The Cragg–Donald F-test statistic 13.22 exceeds this value, so we reject the null hypothesis that the instruments are weak and conclude that they are not weak. This is not the conclusion we would have drawn based on IV/2SLS estimation. The critical value from Table 10A.1 is 22.30, and we would have not rejected the null hypothesis that the instruments are weak.

In Model (4) there are $B = 2$ endogenous variables and $L = 3$ instruments. Using the maximum size of 10% critical value from Table 11A.1 of 5.44, we reject the null hypothesis that the instruments are weak using the Cragg–Donald F-test statistic of 8.60. If we were using the 2SLS/IV estimator, we would have not rejected the hypothesis that the instruments are weak because the critical value from Table 10A.1 is 13.43.

What is indicated by these examples is that the LIML estimator performs better, at least potentially, in the face of weak instruments. We cannot prove anything based on one result from one sample, which is why we present a Monte Carlo simulation experiment in Appendix 11A.3.

EXAMPLE 11.5 | Testing for Weak Instruments with Fuller-Modified LIML

Using the Fuller modification of LIML, and setting the constant $a = 1$, we obtain the estimates in Table 11A.4. All the results are at least somewhat different from the 2SLS/IV estimations, because even for just-identified equations, the Fuller estimator is different from the 2SLS estimator. The only extremely dramatic change now comes in Model (3), where coefficient signs become more in line with the other models, although still nothing is significant. In Model (4), if we adopt the criterion of 10% maximum relative bias, then the Stock–Yogo critical value is 8.96. The Cragg–Donald F-test statistic is 8.6, so we fail to reject the null hypothesis that the instruments are weak.

TABLE 11A.4 **Fuller ($a = 1$) Estimations**

MODEL	(1)	(2)	(3)	(4)
C	17108.0110	15924.1895	2817.5400	18156.7850
	(5.60)	(5.44)	(0.20)	(5.10)
MTR	−18089.5451	−16713.2345	−1304.8205	−18730.1617
	(−5.11)	(−4.93)	(−0.08)	(−4.84)
EDUC	−142.5409	−132.2218	−29.6043	−191.1248
	(−4.41)	(−4.27)	(−0.20)	(−3.05)
KIDSL6	141.4113	105.3703	−287.7915	193.2295
	(1.02)	(0.79)	(−0.65)	(1.21)
NWIFEINC	−101.9491	−94.6401	−12.0108	−102.6290
	(−5.31)	(−5.14)	(−0.15)	(−5.12)
N	428	428	428	428
k	0.9976	1.0172	0.9976	1.0005
FULLER *a*	1.0000	1.0000	1.0000	1.0000
NUMBER IV *L*	1	3	2	3
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60
NUMBER ENDOG *B*	1	1	2	2

t-statistics in parentheses

11A.3 Monte Carlo Simulation Results

In Appendix 10B.2, we carried out a Monte Carlo simulation to explore the properties of the IV/2SLS estimators. Here we employ the same experiment, adding aspects of the new estimators we have introduced in this appendix.

First, examine the percentage rejections of the true null hypothesis $\beta_2 = 1$ using a two-tail test at the 5% level of significance. The Monte Carlo rejection rate for the IV/2SLS estimator is in the column labeled $t\left(\hat{\beta}_2\right)$, and for the LIML estimator in the column $t\left(\hat{\beta}_{2,\text{LIML}}\right)$. The largest difference is in the case of strong endogeneity with weak instruments, in which the test based upon the two-stage least squares estimator rejects 28.86% of the time, while the test based on the LIML estimator rejects 13.47% of the time. Recall that a two-tail test at the 5% level of significance corresponds to determining whether the 95% interval estimate contains the hypothesized parameter value. In these Monte Carlo experiments, the 95% interval estimate based on the LIML estimator contains the true parameter 86.53% of the time, whereas the 95% interval estimate using IV/2SLS contains the true parameter only 71.14% of the time. This finding is consistent with Stock and Yogo's conclusion about coverage rates of the two interval estimation approaches.

In these experiments, there is little difference between the averages of the two-stage least squares estimates, $\bar{\hat{\beta}}_2$ and the Fuller modified ($a = 1$) LIML estimates $\bar{\hat{\beta}}_{2,\text{F}}$. A greater contrast shows up when comparing how close the estimates are to the true parameter value using the mean square error criterion. In Table 11A.5, we report the empirical mean square error for the IV/2SLS estimator, $\text{mse}\left(\hat{\beta}_2\right)$ and that for the Fuller modification of LIML with $a = 4$, $\text{mse}\left(\hat{\beta}_{2,\text{F}}\right)$. Recall that the mean square error measures how close the estimates are to the true parameter value. For the IV/2SLS estimator, the empirical mean square error is

$$\text{mse}\left(\hat{\beta}_2\right) = \sum_{m=1}^{10000}\left(\hat{\beta}_{2m} - \beta_2\right)^2 \Big/ 10{,}000$$

The Fuller-modified LIML has lower mean square error than the IV/2SLS estimator in each experiment, and when the instruments are weak, the improvement is large.

TABLE 11A.5 Monte Carlo Simulation Results

ρ	π	$\bar{F}$	$\bar{\hat{\beta}}_2$	$t\left(\hat{\beta}_2\right)$	$t\left(\hat{\beta}_{2,\text{LIML}}\right)$	$\bar{\hat{\beta}}_{2,\text{F}}$	$\text{mse}\left(\hat{\beta}_2\right)$	$\text{mse}\left(\hat{\beta}_{2,\text{F}}\right)$
0.0	0.1	1.98	0.9941	0.0049	0.0049	0.9941	0.4068	0.0748
0.0	0.5	21.17	0.9998	0.0441	0.0473	0.9997	0.0140	0.0132
0.8	0.1	2.00	1.3311	0.2886	0.1347	1.3375	1.0088	0.3289
0.8	0.5	21.18	1.0111	0.0636	0.0509	1.0000	0.0139	0.0127

CHAPTER 12

Regression with Time-Series Data: Nonstationary Variables

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the differences between stationary and nonstationary time-series processes.
2. Describe the general behavior of an autoregressive process and a random walk process.
3. Explain why we need ''unit root'' tests, and state implications of the null and alternative hypotheses.
4. Explain what is meant by the statement that a series is ''integrated of order one'' or I(1).
5. Perform Dickey–Fuller and augmented Dickey–Fuller tests for stationarity.
6. Explain the meaning of a ''spurious regression.''
7. Explain the concept of cointegration and test whether two series are cointegrated.
8. Explain how to choose an appropriate model for regression analysis with time-series data.

KEYWORDS

autoregressive process
cointegration
Dickey–Fuller test
difference stationary
mean reversion
nonstationary
order of integration
random walks
random walk with drift
spurious regressions
stationary
stochastic process
stochastic trend
tau statistic
trend stationary
unit root tests

The analysis of time-series data is of vital interest to many groups, such as macroeconomists studying the behavior of national and international economies, finance economists analyzing the stock market, and agricultural economists predicting supplies and demands for agricultural products. For example, if we are interested in forecasting the growth of gross domestic product or

inflation, we look at various indicators of economic performance and consider their behavior over recent years. Alternatively, if we are interested in a particular business, we analyze the history of the industry in an attempt to predict potential sales. In each of these cases, we are analyzing time-series data.

We worked with time-series data in Chapter 9 and discovered how regression models for these data often have special characteristics designed to capture their dynamic nature. We saw how including lagged values of the dependent variable or explanatory variables as regressors, or considering lags in the errors, can be used to model dynamic relationships. We showed how autoregressive distributed lag (ARDL) models can be used for forecasting and for computing dynamic multpliers. An important assumption that was maintained throughout Chapter 9 was that the variables are **stationary** and **weakly dependent**. They have means and variances that do not change over time, and autocorrelations that depend on the time between observations, not on the actual time of the observation. Also, their autocorrelations die out, eventually becoming negligible, as the distance between observations increases. There are, however, many economic time series that are not stationary—their means and/or variances change over time—and which exhibit strong dependence—their autocorrelations do not die out or they decline very slowly. In this chapter, we investigate the nature of **nonstationary** variables, examine the consequences of using them in regression analysis, introduce tests for stationarity, and learn how to model regression relationships that involve nonstationary variables. One important new concept that we encounter and which has a bearing on our choice of a regression model is **cointegration**. The widespread use of cointegration and its relevance for many economic time series led to a joint award of the 2003 Nobel Prize in Economics to its developer Clive W.J. Granger.[1]

12.1 Stationary and Nonstationary Variables

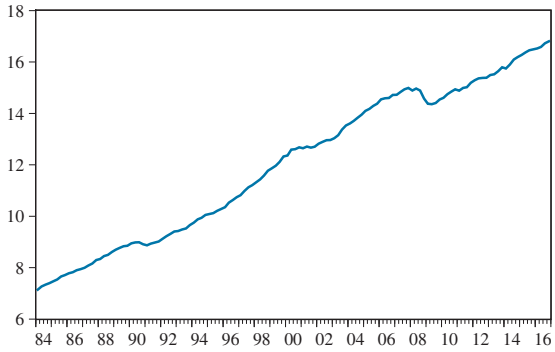
To illustrate the characteristics of nonstationary variables and appreciate their widespread relevance, we begin by examining some important economic variables for the U.S. economy.

EXAMPLE 12.1 | Plots of Some U.S. Economic Time Series

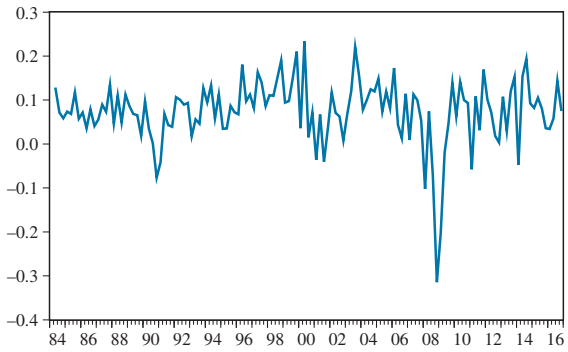
On the left-hand side of Figure 12.1, we display plots of real gross domestic product (a measure of aggregate economic production), the annual inflation rate (*INF*) (a measure of changes in the aggregate price level), the federal funds rate (*FFR*) (the interest rate on overnight loans between banks), and the three-year bond rate (*BR*) (interest rate on a financial asset to be held for three years). The data on gross domestic product (GDP) are quarterly from 1984Q1 to 2016Q4; they can be found in the data file *gdp5*. The data on inflation and the two interest rates are monthly from 1954M8 to 2016M12; they are stored in the data file *usdata5*. *FFR* and *BR* are used for several examples later in the Chapter. Observe how the GDP variable displays upward trending behavior, while the other series "wander up and down" with no discernable pattern or trend.

The figures on the right-hand side of Figure 12.1 are the changes of the corresponding variables on the left-hand side. Recall that we used changes in variables for several of our examples and exercises in Chapter 9. The change in a variable is a particularly important concept used repeatedly in this chapter; it is worth dwelling on its definition. The change in a variable y_t, also known as its **first difference**, is given by $\Delta y_t = y_t - y_{t-1}$. It is the change in the value of the variable y from period $t-1$ to period t. The time series of the changes on the right-hand side of Figure 12.1 display behavior that can be described as irregular ups and downs or more like fluctuations. Changes in the inflation rate and the two interest rates appear to fluctuate around a constant value, approximately zero. Changes in the GDP variable appear to fluctuate around a nonzero value, with a big dip at the time of the global financial crisis. The first question we address in this chapter is: Which data series represent stationary variables and which are observations on nonstationary variables?

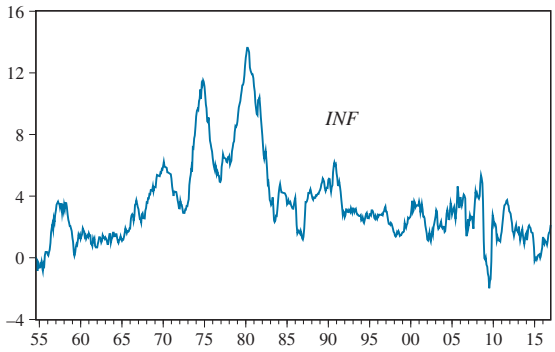
[1] See https://www.britannica.com/biography/Clive-Granger. The corecipient of the 2003 Nobel Prize in Economics was Robert F. Engle whose contribution we consider in Chapter 14.



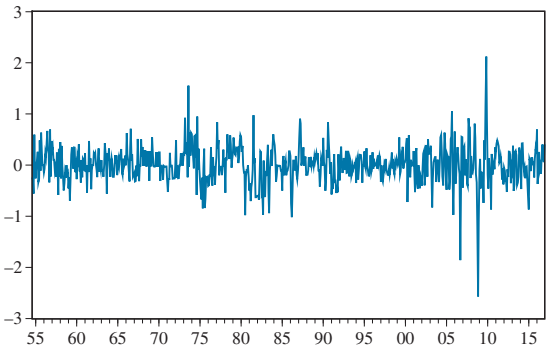
(*a*) Real gross domestic product (GDP), $trillion (quarterly)



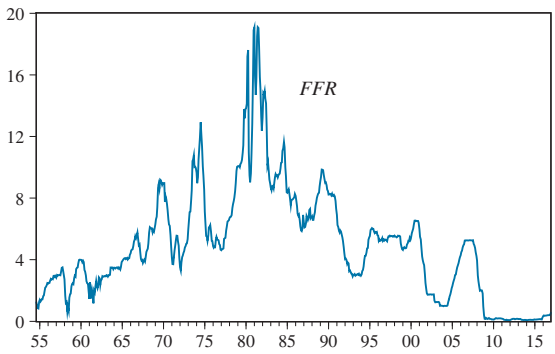
(*b*) Change in GDP (quarterly)



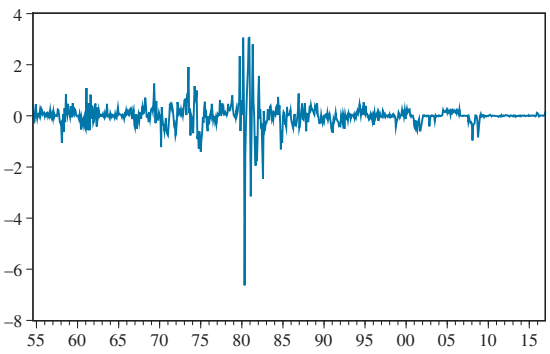
(*c*) Inflation, percent (monthly)



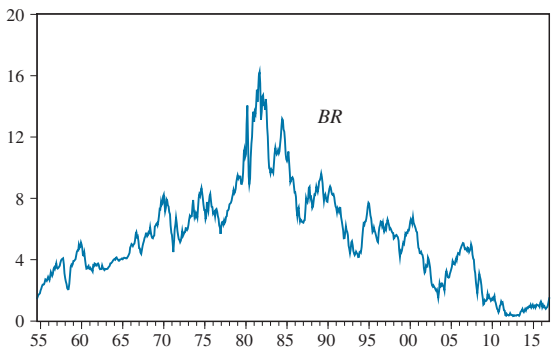
(*d*) Change in inflation (monthly)



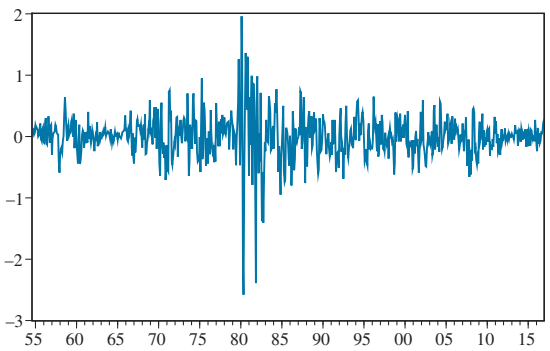
(*e*) Federal funds rate, percent (monthly)



(*f*) Change in the federal funds rate (monthly)



(*g*) Three-year bond rate, percent (monthly)



(*h*) Change in the bond rate (monthly)

FIGURE 12.1 **U.S. Economic Time Series.**

Recall that a stationary time series y_t has mean and variance that are constant over time, and that the covariance (and autocorrelations) between two values from the series depends only on the length of time separating the two values, and not on the actual times at which the values are observed, that is,

$$E(y_t) = \mu \quad \text{(constant mean)} \tag{12.1a}$$

$$\text{var}(y_t) = \sigma^2 \quad \text{(constant variance)} \tag{12.1b}$$

$$\text{cov}(y_t, y_{t+s}) = \text{cov}(y_t, y_{t-s}) = \gamma_s \quad \text{(covariance depends on } s\text{, not } t\text{)} \tag{12.1c}$$

Let us focus on the first condition, that of a constant mean. To investigate whether the means of the series in Figure 12.1 change over time, we divide the observations into two approximately equal subsamples, and compute the sample means for each of these subsamples. They are reported in Table 12.1. Examining the entries in this Table, as well as the plots in Figure 12.1, it is clear that the means of the variables expressed in terms of their original levels do change over time. In Figure 12.1(a), GDP exhibits a clear trend upward leading to a larger mean in the second half of the sample. The other three variables (Figures 12.1(c), (e), and (g)) wander up and then down, making the sample means very sensitive to the period selected. When the sample is divided into two equal parts, more large values appear in the first half of the sample, making the means in that half larger than those in the second half. These characteristics are typical of nonstationary variables. On the other hand, the first differences of the variables (their changes) in Figures 12.1(b), (d), (f), and (h) do not exhibit obvious trends. Their means for the two subsamples are similar in magnitude, particularly when viewed relative to magnitude of their quarter-to-quarter fluctuations. Having a constant mean and fluctuations in the series that tend to return to the mean are characteristics of stationary variables. They have the property of **mean reversion**.

Another characteristic of nonstationary variables is that their sample autocorrelations remain large at long lags. Stationary weakly dependent series have autocorrelations that cut off or tend to decline geometrically, dying out at long lags. The sample autocorrelations of nonstationary series exhibit **strong dependence**. They decline linearly rather than geometrically and are still significant at long lags. As an example, in Figure 12.2, the correlograms for GDP and its change are displayed. The autocorrelations for GDP decline very slowly and continue to be significant, well above the 5% significance bound of 0.17, even at lag 24, an indication that GDP is nonstationary. On the other hand, for the change in GDP, only the first two autocorrelations are significant before the remainder become negligible, suggesting that ΔGDP is stationary.

TABLE 12.1 **Sample Means of Time Series Shown in Figure 12.1**

	Sample Periods		
Variable	GDP *INF*, *BR*, *FFR*	1948Q2 to 2000Q3 1954M8 to 1985M10	2000Q4 to 2016Q4 1985M11 to 2016M12
Real GDP (a)		9.56	14.68
Inflation rate (c)		4.42	2.59
Federal funds rate (e)		6.20	3.65
Bond rate (g)		6.56	4.29
Change in GDP (b)		0.083	0.065
Change in the inflation rate (d)		0.01	−0.003
Change in the federal funds rate (f)		0.02	−0.02
Change in the bond rate (h)		0.02	−0.02

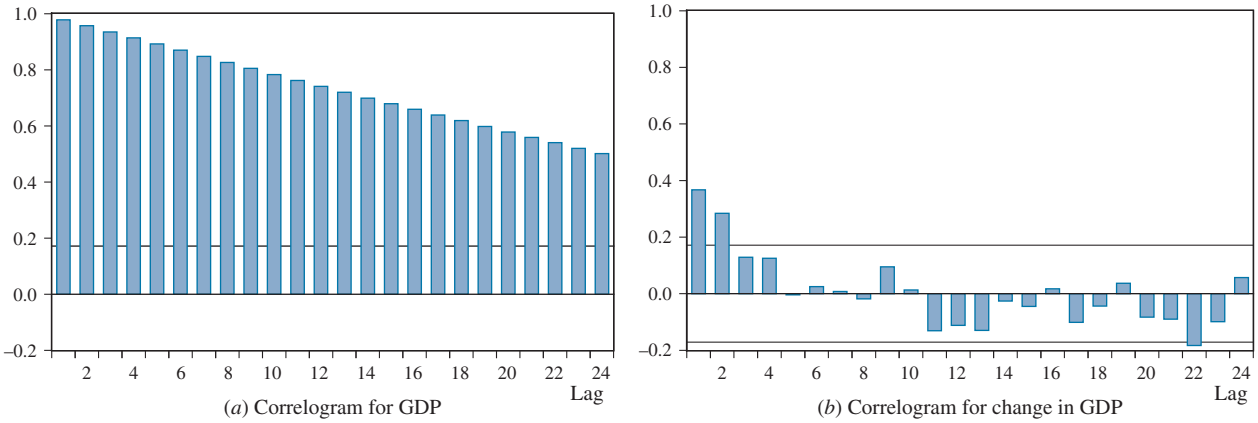


FIGURE 12.2 **Correlograms for GDP and the change in GDP.**

Plotting a series, examining whether its mean changes over time, and checking its sample autocorrelations give some indication of whether a series is stationary or nonstationary, but these checks are not conclusive, and they lack the rigor of a formal hypothesis test. Also, our discovery that series with nonstationary characteristics have stationary characteristics after first differencing is a common occurrence, but it is not universal and it needs verification. Formal testing for stationarity is introduced in Section 12.3. Before then, we discuss modeling series with trends and the consequences of nonstationarity for least-squares regressions.

12.1.1 Trend Stationary Variables

In Example 12.1, we saw how GDP has a definite trend, making it nonstationary, and that the other variables—inflation and the two interest rates—tend to wander up and down, another characteristic of nonstationary variables. Nonstationary variables that wander up and down, trending in one direction and then the other, are said to possess a **stochastic trend**. Definite trends, upward or downward, can be attributable to a stochastic trend or a **deterministic trend**, and sometimes both. Variables that are stationary after "subtracting out" a deterministic trend are called **trend stationary**. In this Section, we consider the notion of a deterministic trend, how it relates to the concept of trend stationarity, and the modeling of regression relationships involving trend stationary variables. Stochastic trends are introduced in Section 12.1.3.

The simplest model for a deterministic trend for a variable y is the linear trend model

$$y_t = c_1 + c_2 t + u_t \tag{12.2}$$

where $t = 1, 2, \ldots, T$. If we focus just on the trend and assume any change in the error is zero $(\Delta u_t = u_t - u_{t-1} = 0)$, then the coefficient c_2 gives the change in y from a one period to the next

$$y_t - y_{t-1} = (c_1 + c_2 t) - [c_1 + c_2 (t-1)] + \Delta u_t = c_2$$

The "time variable" t does not necessarily have to start at "1" and increase in increments of "1". Redefining it using a linear transformation, say $t^* = a + bt$, simply changes the values for c_1 and c_2 and changes the interpretation of c_2 if $b \neq 0$. The trend $c_1 + c_2 t$ is called a deterministic trend because it does not contain a stochastic (random) component. The variable y_t is trend stationary if its fluctuations around this trend are stationary. Since these fluctuations are given by changes in the error term

$$u_t = y_t - (c_1 + c_2 t) \tag{12.3}$$

y_t is trend stationary if u_t is stationary.

When y_t is trend stationary, we can use least squares to find estimates $\hat{c}_1$ and $\hat{c}_2$ from (12.2) and then convert the trend stationary variable y_t to a stationary variable $\hat{u}_t$ by removing the trend:

$$\hat{u}_t = y_t - (\hat{c}_1 + \hat{c}_2 t) \tag{12.4}$$

If we are considering a regression or an ARDL model involving two trend stationary variables, say y_t and x_t, then, after their trends have been removed, making them stationary, their relationship can be estimated within the framework of Chapter 9.

To explore this notion further, suppose $y_t = c_1 + c_2 t + u_t$ and $x_t = d_1 + d_2 t + v_t$ are trend stationary variables; both u_t and v_t are stationary. To estimate a relationship between y_t and x_t, we first remove their trends: $\tilde{y}_t = y_t - (\hat{c}_1 + \hat{c}_2 t)$ and $\tilde{x}_t = x_t - (\hat{d}_1 + \hat{d}_2 t)$ where $\hat{c}_1, \hat{c}_2, \hat{d}_1$ and $\hat{d}_2$ are the least-squares estimates from the respective trends. We have used the notation $\tilde{y}_t$ and $\tilde{x}_t$ instead of $\hat{u}_t$ and $\hat{v}_t$ in line with that used in the FWL theorem introduced in Section 5.2.4. If we hypothesize that changes in y around its trend are related to changes in x around its trend, without any lags, a suitable linear model is

$$\tilde{y}_t = \beta \tilde{x}_t + e_t \tag{12.5}$$

An intercept can be omitted because $\tilde{y}_t$ and $\tilde{x}_t$ are OLS residuals with zero means. Now, we know from the FWL theorem that the OLS estimate of β from (12.5) is identical to the OLS estimate of β from the equation

$$y_t = \alpha_1 + \alpha_2 t + \beta x_t + e_t \tag{12.6}$$

Thus, when y and x are trend stationary, we can estimate a relationship between them by first removing the trends or by including a trend variable in the equation.

With trend stationary variables in more general ARDL models, we can proceed in a similar way, estimating either

$$\tilde{y}_t = \sum_{s=1}^{p} \theta_s \tilde{y}_{t-s} + \sum_{r=0}^{q} \delta_r \tilde{x}_{t-r} + e_t \tag{12.7}$$

or

$$y_t = \alpha_1 + \alpha_2 t + \sum_{s=1}^{p} \theta_s y_{t-s} + \sum_{r=0}^{q} \delta_r x_{t-r} + e_t \tag{12.8}$$

Assuming we create $\tilde{y}_{t-s}$ and $\tilde{x}_{t-r}$ by lagging $\tilde{y}_t$ and $\tilde{x}_t$, not by separately detrending every lag of y and x, there will be some slight differences in the estimates from (12.7) and (12.8).

Because trend stationary variables do not introduce any special problems providing a trend is included or the variables are detrended, they are often simply referred to as "stationary," although, strictly speaking, they are not stationary because their means change over time. Also, it is important not to ignore any trend. Estimating the model $y_t = \alpha_1 + \beta x_t + e_t$ when both y_t and x_t have deterministic trends can suggest a significant relationship between y_t and x_t even when none exists.

It is useful to pause at this point to emphasize what we have established and what we have not yet covered. We have discovered that regression relationships between trend stationary variables can be modeled by removing the deterministic trend from the variables, making them stationary, or by including the deterministic trend directly in the equation. What we have not yet covered is how to distinguish between deterministic trends and stochastic trends and how to model regression relationships between nonstationary variables with stochastic trends. In Example 12.1, GDP had an obvious trend. We do not yet know whether this trend is deterministic or stochastic, or how it should be modeled within a regression framework. We address these questions in the upcoming sections, but first it is useful to note that the linear trend in (12.2) is not the only possible deterministic trend, and to give an example.

Other Trends Another popular trend is one where, on average, a variable is growing at a constant *percentage* rate. If we momentarily ignore the error term, then, for a proportional change a_2, we have $y_t = y_{t-1} + a_2 y_{t-1}$, or, in percentage terms,

$$100 \times \left(\frac{y_t - y_{t-1}}{y_{t-1}} \right) = 100 a_2$$

Recognizing that $(y_t - y_{t-1})/y_{t-1}$ can be approximated by $\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$, we have

$$\ln(y_t) - \ln(y_{t-1}) \cong \%\Delta y_t = 100 a_2$$

A model with this property, with an error term included, is

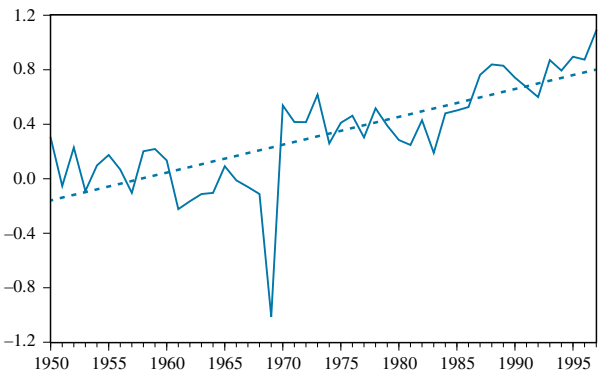
$$\ln(y_t) = a_1 + a_2 t + u_t \tag{12.9}$$

In this case. the deterministic trend for y_t is $\exp(a_1 + a_2 t)$, and $\ln(y_t)$ will be trend stationary if u_t is stationary. This model was introduced earlier in Section 4.5.1 in the context of modeling increases in wheat yield that are attributable to technological change. It may pay to go back and reread that Section now; it will give you more insights into the constant growth rate model.

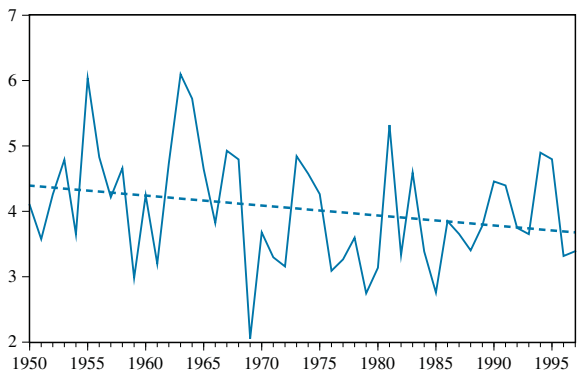
The deterministic trend models in (12.2) and (12.9) are the most common, but others are possible. In Section 4.4.2, the cubic trend $y_t = \beta_1 + \beta_2 t^3 + e_t$ was used to model wheat yield. In Exercises 5.21 and 5.22, the interaction variable $TREND \times RAIN$ was included. A quadratic trend was used to model a decreasing and then increasing income share in Exercises 6.28 and 6.29. However, most deterministic trends tend to be continuously increasing or decreasing in which case quadratic or cubic trends that eventually turn up or down may not be well suited. A restricted range of the curve may fit the data well for the sample period, but outside this range a quadratic or cubic may be unrealistic. For this reason, the deterministic trends implied by (12.2) and (12.9) are the most popular.

EXAMPLE 12.2 | A Deterministic Trend for Wheat Yield

Scientists are continually working on ways to increase global food production to keep pace with a growing world population. One small contribution to this effort is the work of agronomists who develop new varieties of wheat to increase wheat yield. In the Toodyay Shire of Western Australia, we expect wheat yield to be trending upward over time reflecting the development of new varieties. However, wheat growing in Western Australia is a risky business. Its success depends heavily on rainfall, which is not always reliable. Thus, we expect yield to fluctuate around an increasing trend. Data on annual wheat yield and rainfall during the growing season for the Toodyay Shire, from 1950 to 1997, can be found in the data file *toody5*. For wheat yield, we use the constant growth rate trend $\ln(YIELD_t) = a_1 + a_2 t + u_t$. The observations for $\ln(YIELD_t)$ are plotted in Figure 12.3(a), along with the linear trend line. The observations fluctuate around the increasing trend with a particularly bad year in 1969. Examining the rainfall data in Figure 12.3(b), we



(*a*) ln(*YIELD*)



(*b*) *RAIN*

FIGURE 12.3 **Plots of time series for wheat yield example.**

discover there is a slight downward trend and very little rainfall in 1969.

It turns out that there are decreasing returns to rainfall and so we include $RAIN^2$ as well as $RAIN$ in the model, leading to the following estimated equation

$$\underset{(\text{se})}{\ln(YIELD_t)} = \underset{(0.00252)}{-2.510 + 0.01971t} + \underset{(0.290)}{1.149RAIN_t} \underset{(0.0346)}{- 0.1344RAIN_t^2} + \hat{e}_t \tag{12.10}$$

The other alternative is to detrend ln(*YIELD*), *RAIN*, and $RAIN^2$ and to estimate the detrended model. First, estimating the trends, we obtain

$$\underset{(\text{se})}{\widehat{\ln(YIELD_t)}} = \underset{(0.00276)}{-0.1801 + 0.02044t}$$

$$\underset{(\text{se})}{\widehat{RAIN_t}} = \underset{(0.00891)}{4.408 - 0.01522t}$$

$$\underset{(\text{se})}{\widehat{RAIN_t^2}} = \underset{(0.0747)}{20.35 - 0.1356t}$$

The first two equations describe the trend lines in Figure 12.3. After computing $RRAIN_t = RAIN_t - \widehat{RAIN_t}$, $RRAIN2_t = RAIN_t^2 - \widehat{RAIN_t^2}$, and $RLYIELD_t = \ln(YIELD_t) - \widehat{\ln(YIELD_t)}$, we obtain

$$\underset{(\text{se})}{\widehat{RLYIELD_t}} = \underset{(0.284)}{1.149RRAIN_t} \underset{(0.0339)}{- 0.1344RRAIN2_t} \tag{12.11}$$

Notice the estimates in (12.10) and (12.11) are identical, but the standard errors are not. The standard error discrepancy arises from the different degrees of freedom used to estimate the error variance. In (12.10), it is $48 - 4 = 44$; in (12.11), it is $48 - 2 = 46$. We can correct the standard errors in (12.11) by multiplying them by $\sqrt{46/44} = 1.022$. In large samples, the difference will be negligible. The legitimacy of the estimates in (12.10) and (12.11) depends on the assumption that ln(*YIELD*), *RAIN*, and $RAIN^2$ are trend stationary. This assumption can be checked using the hypothesis testing machinery that is developed in Section 12.3 (see Exercise 12.16).

12.1.2 The First-Order Autoregressive Model

To develop a framework for modeling nonstationary variables that possess a stochastic trend, we begin by revising the first-order autoregressive AR(1) model that was introduced in Chapter 9.

The econometric model generating a time-series variable y_t is called a **stochastic** or **random process**. A sample of observed y_t values is called a particular **realization** of the **stochastic process**. It is one of many possible paths that the stochastic process could have taken. Univariate time-series models are examples of stochastic processes where a single variable y is related to past values of itself and current and past error terms. In contrast to regression modeling, univariate time-series models do not contain any explanatory variables (no x's).

The AR(1) model is a useful univariate time-series model for explaining the difference between stationary and nonstationary series. We first consider an AR(1) model with a zero mean given by

$$y_t = \rho y_{t-1} + v_t, \quad |\rho| < 1 \tag{12.12}$$

where the errors v_t are independent, with zero mean and constant variance σ_v^2, and may be normally distributed. In the context of time-series models, the errors are sometimes known as "shocks" or "innovations." As we will see, the assumption $|\rho| < 1$ implies that y_t is stationary. The AR(1) process shows that each realization of the random variable y_t contains a proportion ρ of last period's value y_{t-1} plus an error v_t drawn from a distribution with mean zero and variance σ_v^2. Since we are concerned with only one lag, the model is described as an autoregressive model of order one. In general, an AR(p) model includes lags of the variable y_t up to y_{t-p}. An example of an AR(1) time series with $\rho = 0.7$ and independent $N(0, 1)$ random errors is shown in Figure 12.4a. Note that the data have been artificially generated. Observe how the time series fluctuates around zero and has no trend-like behavior, a characteristic of stationary series.

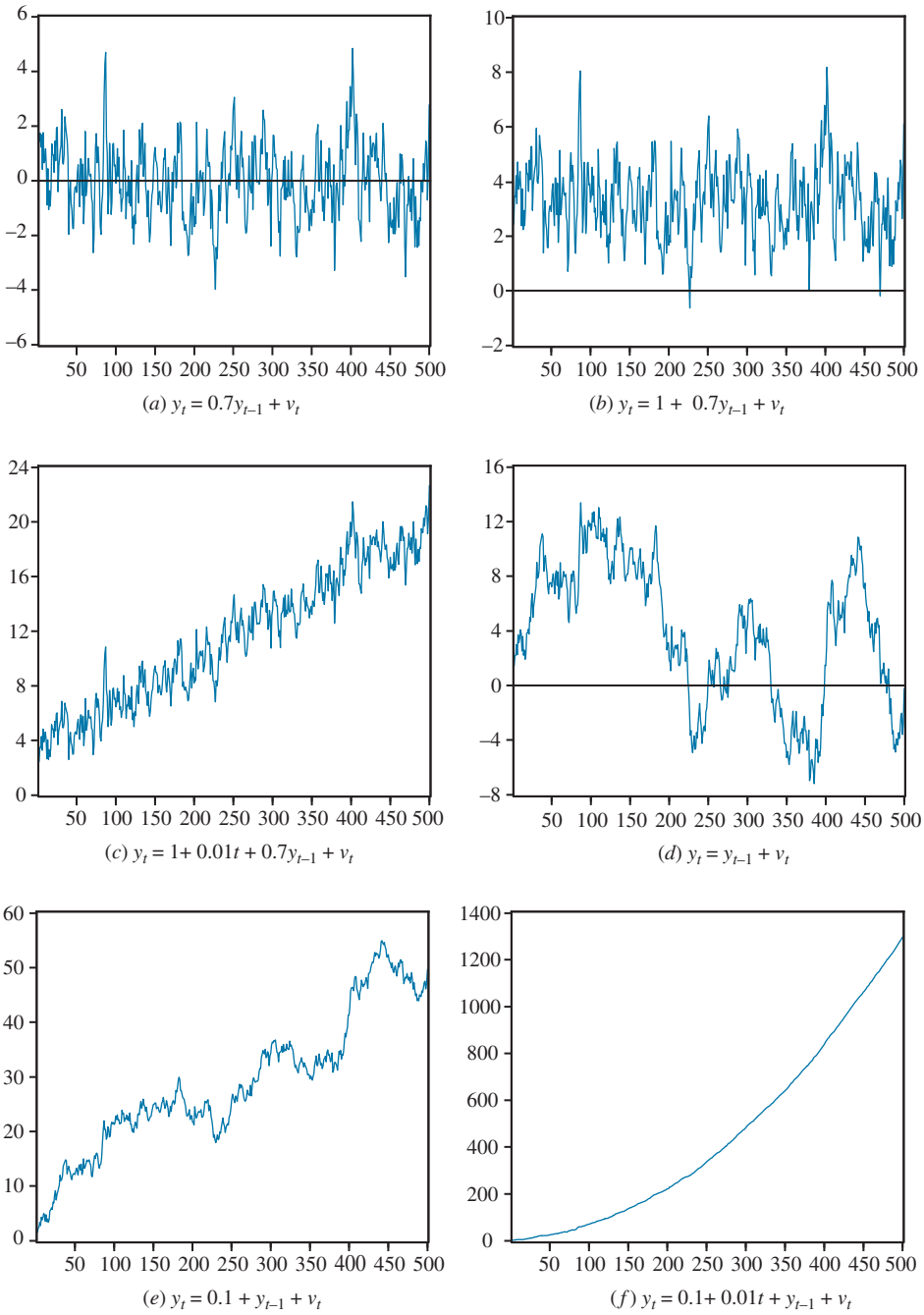


FIGURE 12.4 **Time-series models.**

The value "zero" is the constant mean of the series, and it can be determined by doing some algebra known as **recursive substitution**.[2] Consider the value of y at time $t = 1$, then its value at time $t = 2$, and so on. These values are

[2]An alternative to recursive substitution when the variable is stationary is to use the lag operator algebra discussed in Section 9.5.4.

$$
\begin{aligned}
y_1 &= \rho y_1 + v_1 \\
y_2 &= \rho y_1 + v_2 = \rho\left(\rho y_0 + v_1\right) + v_2 = \rho^2 y_0 + \rho v_1 + v_2 \\
&\vdots \\
y_t &= v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots + \rho^t y_0
\end{aligned}
$$

The mean of y_t is

$$E\left(y_t\right) = E\left(v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots\right) = 0$$

since the error v_t has zero mean, and the value of $\rho^t y_0$ is negligible for a large t. In Appendix 9B, the variance was shown to be a constant $\sigma_v^2/\left(1-\rho^2\right)$, while the covariance between two errors s periods apart γ_s is $\sigma_v^2\rho^s/\left(1-\rho^2\right)$. Thus, the AR(1) model in (12.12) is a classic example of a stationary process with a zero mean.

Real-world data rarely have a zero mean. We can introduce a nonzero mean μ by replacing y_t in (12.12) with $\left(y_t - \mu\right)$ as follows:

$$\left(y_t - \mu\right) = \rho\left(y_{t-1} - \mu\right) + v_t$$

which can then be rearranged as

$$y_t = \alpha + \rho y_{t-1} + v_t, \quad |\rho| < 1 \tag{12.13}$$

where $\alpha = \mu(1-\rho)$, that is, we can accommodate a nonzero mean in y_t by either working with the "demeaned" variable $\left(y_t - \mu\right)$ or introducing the intercept term α in the **autoregressive process** of y_t as in (12.13). Corresponding to these two ways, we describe the "de-meaned" variable $\left(y_t - \mu\right)$ as being stationary around zero, or the variable y_t as stationary around its mean value $\mu = \alpha/(1-\rho)$.

An example of a time series that follows this model, with $\alpha = 1$, $\rho = 0.7$ is shown in Figure 12.4(b). We have used the same values of the error v_t as in Figure 12.4(a), so the figure shows the added influence of the constant term. Note that the series now fluctuates around a nonzero value. This nonzero value is the constant mean of the series

$$E\left(y_t\right) = \mu = \alpha/(1-\rho) = 1/(1-0.7) = 3.33$$

Another extension to (12.12) is to consider an AR(1) model fluctuating around a linear trend $(\mu + \delta t)$. In this case, we let the "detrended" series $(y_t - \mu - \delta t)$ behave like an autoregressive model

$$\left(y_t - \mu - \delta t\right) = \rho\left[y_{t-1} - \mu - \delta(t-1)\right] + v_t, \quad |\rho| < 1$$

which can be rearranged as

$$y_t = \alpha + \rho y_{t-1} + \lambda t + v_t \tag{12.14}$$

where $\alpha = \left[\mu(1-\rho) + \rho\delta\right]$ and $\lambda = \delta(1-\rho)$. For $|\rho| < 1$, equation (12.14) is an example of a trend-stationary process. Figure 12.4(c) displays a plot of this process for parameters $\rho = 0.7$, $\alpha = 1$, and $\lambda = 0.01$. The detrended series $\left(y_t - \mu - \delta t\right)$ has a constant variance, and covariances that depend only on the time separating observations, not the time at which they are observed. In other words, the detrended series is stationary; y_t is stationary around the deterministic trend line $\mu + \delta t$.

12.1.3 Random Walk Models

Consider the special case of $\rho = 1$ in (12.12):

$$y_t = y_{t-1} + v_t \tag{12.15}$$

This model is known as the random walk model. Equation (12.15) shows that each realization of the random variable y_t contains last period's value y_{t-1} plus an error v_t. An example of a time series that can be described by this model is shown in Figure 12.4(d). These time series are called **random walks** because they appear to wander slowly upward or downward with no real pattern; the values of sample means calculated from subsamples of observations will be dependent on the sample period, a characteristic of nonstationary series.

We can understand the "wandering" behavior of random walk models by doing some recursive substitution.

$$\begin{aligned} y_1 &= y_0 + v_1 \\ y_2 &= y_1 + v_2 = (y_0 + v_1) + v_2 = y_0 + \sum_{s=1}^{2} v_s \\ &\vdots \\ y_t &= y_{t-1} + v_t = y_0 + \sum_{s=1}^{t} v_s \end{aligned}$$

The random walk model contains an initial value y_0 (often set to zero because it is so far in the past that its contribution to y_t is negligible) plus a component that is the sum of the past stochastic terms $\sum_{s=1}^{t} v_s$. This latter component is called the **stochastic trend**. This term arises because a stochastic component v_t is added for each time t, and because it causes the time series to trend in unpredictable directions. If the variable y_t is subjected to a sequence of positive shocks, $v_t > 0$, followed by a sequence of negative shocks, $v_t < 0$, it will have the appearance of wandering upward, then downward.

We have used the fact that y_t is a sum of errors to explain graphically the nonstationary nature of the random walk. We can also use it to show algebraically that the conditions for stationarity do not hold. Recognizing that the v_t are independent with zero means and identical variances σ_v^2, taking the expectation and the variance of y_t yields, for a fixed initial value y_0,

$$\begin{aligned} E(y_t) &= y_0 + E(v_1 + v_2 + \cdots + v_t) = y_0 \\ \text{var}(y_t) &= \text{var}(v_1 + v_2 + \cdots + v_t) = t\sigma_v^2 \end{aligned}$$

The random walk has a mean equal to its initial value and a variance that increases over time, eventually becoming infinite. Although the mean is constant, the increasing variance implies that the series may not return to its mean, and so sample means taken for different periods are not the same.

Another nonstationary model is obtained by adding a constant term to (12.15):

$$y_t = \delta + y_{t-1} + v_t \tag{12.16}$$

This model is known as the **random walk with drift**. Equation (12.16) shows that each realization of the random variable y_t contains an intercept (the drift component δ) plus last period's value y_{t-1} plus the error v_t. An example of a time series that can be described by this model (with $\delta = 0.1$) is shown in Figure 12.4(e). Notice how the time-series data appear to be "wandering" as well as "trending" upward. In general, random walk with drift models show definite trends either upward (when the drift δ is positive) or downward (when the drift δ is negative).

Again, we can get a better understanding of this behavior by applying recursive substitution:

$$\begin{aligned} y_1 &= \delta + y_0 + v_1 \\ y_2 &= \delta + y_1 + v_2 = \delta + (\delta + y_0 + v_1) + v_2 = 2\delta + y_0 + \sum_{s=1}^{2} v_s \\ &\vdots \\ y_t &= \delta + y_{t-1} + v_t = t\delta + y_0 + \sum_{s=1}^{t} v_s \end{aligned}$$

The value of y at time t is made up of an initial value y_0, the stochastic trend component $\left(\sum_{s=1}^{t} v_s\right)$, and now a deterministic trend component $t\delta$. It is called a deterministic trend because a fixed value

δ is added for each time t. The variable y wanders up and down as well as increases by a fixed amount at each time t. The mean and variance of y_t are

$$E(y_t) = t\delta + y_0 + E(v_1 + v_2 + v_3 + \cdots + v_t) = t\delta + y_0$$
$$\text{var}(y_t) = \text{var}(v_1 + v_2 + v_3 + \cdots + v_t) = t\sigma_v^2$$

In this case, both the constant mean and constant variance conditions for stationarity are violated.

We can extend the random walk model even further by adding a time trend:

$$y_t = \alpha + \delta t + y_{t-1} + v_t \tag{12.17}$$

An example of a time series that can be described by this model (with $\alpha = 0.1; \delta = 0.01$) is shown in Figure 12.4(f). Note how the addition of a time-trend variable t strengthens the trend behavior. We can see the amplification using the same algebraic manipulation as before:

$$y_1 = \alpha + \delta + y_0 + v_1$$
$$y_2 = \alpha + \delta 2 + y_1 + v_2 = \alpha + 2\delta + (\alpha + \delta + y_0 + v_1) + v_2 = 2\alpha + 3\delta + y_0 + \sum_{s=1}^{2} v_s$$
$$\vdots$$
$$y_t = \alpha + \delta t + y_{t-1} + v_t = t\alpha + \left(\frac{t(t+1)}{2}\right)\delta + y_0 + \sum_{s=1}^{t} v_s$$

where we have used the formula for a sum of an arithmetic progression,

$$1 + 2 + 3 + \cdots + t = t(t+1)/2$$

The additional term has the effect of strengthening the trend behavior.

To recap, we have considered the autoregressive class of models and have shown that they display properties of stationarity when $|\rho| < 1$. We have also discussed the random walk class of models when $\rho = 1$. We showed that random walk models display properties of nonstationarity. Now, go back and compare the real-world data in Figure 12.1 with those in Figure 12.4. Ask yourself what models might have generated the different data series in Figure 12.1. In the next few sections we shall consider how to test which series in Figure 12.1 exhibit properties associated with stationarity, as well as which series exhibit properties associated with nonstationarity.

12.2 Consequences of Stochastic Trends

In Section 12.1.2, we noted that regressions involving variables with a deterministic trend, *and no stochastic trend*, did not present any difficulties providing the trend was included in the regression relationship, or the variables were detrended. Allowing for the trend was important because excluding it could lead to omitted variable bias. Now we consider the implications of estimating regressions involving variables with stochastic trends. In this context, because stochastic trends are the most prevalent source of nonstationarity, and they introduce special problems, when we refer to nonstationary variables, we will generally mean variables that are neither stationary nor trend stationary.

A consequence of proceeding with the regression involving nonstationary variables with stochastic trends is that OLS estimates no longer have approximate normal distributions in large samples. That means interval estimates and hypothesis tests will no longer be valid. Precision of estimation may not be what it seems to be and conclusions about relationships between variables could be wrong. One particular hazard is that two totally independent random walks can appear to have a strong linear relationship when none exists. Outcomes of this nature have been given the name **spurious regressions**.

EXAMPLE 12.3 | A Regression with Two Random Walks

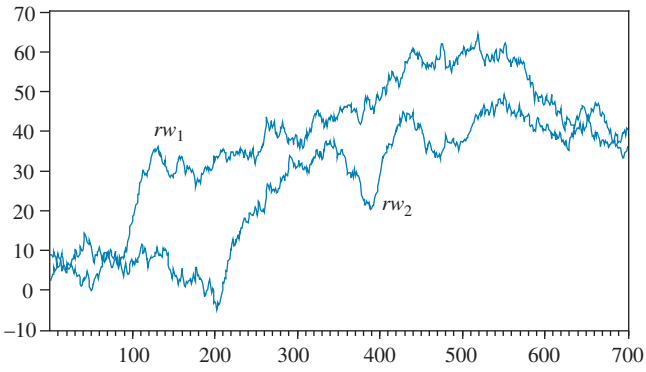
To illustrate the spurious regression problem, consider the following two independent random walks:

$$rw_1 : y_t = y_{t-1} + v_{1t}$$
$$rw_2 : x_t = x_{t-1} + v_{2t}$$

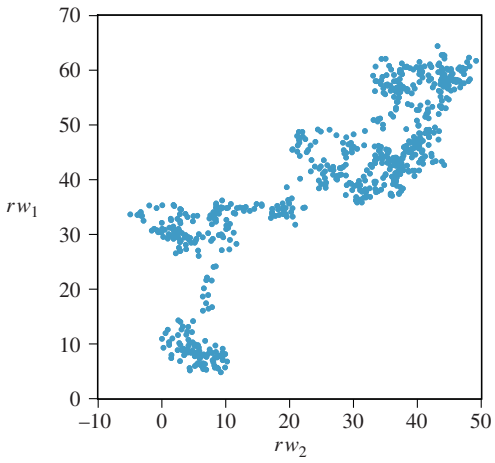
where v_{1t} and v_{2t} are independent $N(0, 1)$ random errors. Two such series are shown in Figure 12.5(a)—the data are in the data file *spurious*. These series were generated independently and, in truth, have no relation to one another, yet when we plot them, as we have done in Figure 12.5(b), we see a positive relationship between them. If we estimate a simple regression of series one (rw_1) on series two (rw_2), we obtain the following results:

$$rw_{1t} = 17.818 + 0.842 rw_{2t}, \quad R^2 = 0.70$$
$$(t) \qquad\qquad (40.837)$$

This result suggests that the simple regression model fits the data well ($R^2 = 0.70$) and that the estimated slope is significantly different from zero. In fact, the t-statistic is huge! These results are, however, completely meaningless, or spurious. The apparent significance of the relationship is false. It results from the fact that we have related one series with a stochastic trend to another series with another stochastic trend. In fact, these series have nothing in common, nor are they causally related in any way. Similar and more dramatic results are obtained when random walk with drift series are used in regressions. Typically the residuals from such regressions will be highly correlated. For this example, the LM test value to test for first-order autocorrelation (p-value in parenthesis) is 682.958 (0.000); a sure sign that there is a problem with the regression.



(*a*) Time series



(*b*) Scatter plot

FIGURE 12.5 **Time series and scatter plot of two random walk variables.**

To summarize, when nonstationary time series are used in a regression model, the results may spuriously indicate a significant relationship when there is none. In these cases the least-squares estimator and least-squares predictor do not have their usual properties, and t-statistics are not reliable. Since many macroeconomic time series are nonstationary, it is particularly important to take care when estimating regressions with macroeconomic variables.

There are also important policy considerations for distinguishing between stationary and nonstationary variables. With nonstationary variables each error or shock v_t has a lasting effect, and these shocks accumulate. With stationary variables the effect of a shock eventually dies out and the variable returns to its mean. Whether a change in a macroeconomic variable has a permanent or transitory effect is essential information for policy makers.

How then can we test whether a series is stationary or nonstationary, and how do we conduct regression analysis with nonstationary data? The former is discussed in Section 12.3, while the latter is considered in Section 12.4.

12.3 Unit Root Tests for Stationarity

There are many tests for assessing whether a series is stationary or nonstationary. The most popular one, and the one that we discuss in detail, is the **Dickey–Fuller test** for a **unit root**. What do we mean by a "unit root"? Because you will hear this term frequently when nonstationary time series are being discussed, it is useful to digress for a moment to explain its origin.

12.3.1 Unit Roots

We have seen that in the AR(1) model $y_t = \alpha + \rho y_{t-1} + v_t$, y_t is stationary if $|\rho| < 1$ and nonstationary if $\rho = 1$. We also say that y_t has a unit root if $\rho = 1$, but to appreciate the origin of the term, we need to consider the more general AR(p) model $y_t = \alpha + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + v_t$. In this model, y_t is stationary if the roots of the polynomial equation

$$\varphi(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_p z^p \tag{12.18}$$

are greater than one in absolute value. The roots are the values of z that satisfy the equation $\varphi(z) = 0$. When $p = 1$ and $y_t = \alpha + \theta_1 y_{t-1} + v_t$, we have $\varphi(z) = 1 - \theta_1 z = 0$, and $z = 1/\theta_1$. The condition for stationarity is $|z| > 1$, which is the same as $|\theta_1| < 1$. If, in (12.18), one of the roots is equal to one, then y_t is said to have a unit root. It has a stochastic trend and is nonstationary. When $p = 1$ and $\varphi(z) = 1 - \theta_1 z = 0$, then $z = 1$ implies $\theta_1 = 1$. Note that we have used θ_1 and ρ interchangeably for the AR(1) model. It is convenient to use θ_1 when considering the AR(1) process as a special case of an AR(p) process. Using ρ emphasizes that the coefficient of y_{t-1} in an AR(1) process is the first-order autocorrelation.

To summarize, if y_t has a unit root, it is nonstationary. For y_t to be stationary, the roots of (12.18) must be greater than one in absolute value. In the AR(1) model $y_t = \alpha + \rho y_{t-1} + v_t$, these conditions translate into $\rho = 1$ for the unit root and $|\rho| < 1$ for stationarity. In higher-order AR models, the conditions for a unit root and for stationarity, written in terms of the parameters $\theta_1, \theta_2, \ldots, \theta_p$, are more complicated. We explore these conditions for the AR(2) model in Exercise 12.1.

You might be wondering what happens if one of the roots of $\varphi(z)$ is less than one in absolute value. Or, in particular, what happens if $\rho > 1$ in the AR(1) process. In this case, y_t is nonstationary and explosive. Empirically, we do not observe time series that explode and so we restrict ourselves to unit roots and roots that imply a stationary process. In the Dickey–Fuller tests that follow the null hypothesis is that y_t has a unit root and the alternative is that y_t is stationary.

12.3.2 Dickey–Fuller Tests

There are three variations of the Dickey–Fuller test, each one designed for a different alternative hypothesis.

1. The alternative hypothesis is that y_t is stationary around a nonzero mean. An example of such a series is that depicted in Figure 12.4(b). In this case, the test equation includes an intercept but no trend term.
2. The alternative hypothesis is that y_t is stationary around a linear deterministic trend, like that depicted in Figure 12.4(d). Here, the test equation includes both intercept and trend terms.
3. The alternative hypothesis is that y_t is stationary around a zero mean as illustrated in Figure 12.4(a). Both intercept and trend are excluded from the test equation in this case.

The choice between these tests can be guided by the nature of the data, revealed by plotting the series against time. If it is not obvious from a plot which test is the most relevant—and it will not always be obvious—more than one test equation can be used to check the robustness of a test conclusion.

12.3.3 Dickey–Fuller Test with Intercept and No Trend

Consider a time series y_t that has no definite continuous trend upward or downward, and that is not obviously centered around zero. Suppose we wish to test whether this series is better represented by a stationary AR(1) process like that in Figure 12.4(b) or a nonstationary random walk like that in Figure 12.4(d). The nonstationary random walk is set up as the null hypothesis

$$H_0 : y_t = y_{t-1} + v_t \tag{12.19}$$

and the stationary AR(1) process becomes the alternative hypothesis

$$H_1 : y_t = \alpha + \rho y_{t-1} + v_t \quad |\rho| < 1 \tag{12.20}$$

Throughout, we assume the v_t are independent random errors with mean zero and variance σ_v^2, and that they are uncorrelated with the past values $y_{t-1}, y_{t-2}, \ldots$. Under H_1, the series fluctuates around a constant mean. Under H_0, it wanders upward and downward but does not exhibit a clear trend in either direction and does not tend to return to a constant mean.

An obvious way to specify the null hypothesis in terms of the parameters in the unrestricted alternative is $H_0 : \alpha = 0,\ \rho = 1$. A test for this purpose has been developed,[3] but it has become more common to simply specify the null as $H_0 : \rho = 1$. One way to justify omission of $\alpha = 0$ from H_0 is to recall that $\alpha = \mu(1 - \rho)$. If $\rho = 1$, then $\alpha = 0$, and so one can argue that testing $H_0 : \rho = 1$ is sufficient. Thus, we test for nonstationary in the AR(1) model $y_t = \alpha + \rho y_{t-1} + v_t$, by testing $H_0 : \rho = 1$ against the alternative $H_1 : |\rho| < 1$, or simply $H_1 : \rho < 1$. This one-sided (left tail) test is put into a more convenient form by subtracting y_{t-1} from both sides of (12.20) to obtain:

$$\begin{aligned} y_t - y_{t-1} &= \alpha + \rho y_{t-1} - y_{t-1} + v_t \\ \Delta y_t &= \alpha + (\rho - 1) y_{t-1} + v_t \\ &= \alpha + \gamma y_{t-1} + v_t \end{aligned} \tag{12.21}$$

where $\gamma = \rho - 1$ and $\Delta y_t = y_t - y_{t-1}$. Then, the hypotheses can be written either in terms of ρ or in terms of γ:

$$\begin{aligned} H_0 : \rho = 1 \quad &\Longleftrightarrow \quad H_0 : \gamma = 0 \\ H_1 : \rho < 1 \quad &\Longleftrightarrow \quad H_1 : \gamma < 0 \end{aligned} \tag{12.22}$$

[3] An advanced reference is Hamilton, J.D. (1994), *Time Series Analysis*, Princeton, p. 494.

TABLE 12.2 **Critical Values for the Dickey–Fuller Test**

Model	1%	5%	10%
$\Delta y_t = \gamma y_{t-1} + v_t$	−2.56	−1.94	−1.62
$\Delta y_t = \alpha + \gamma y_{t-1} + v_t$	−3.43	−2.86	−2.57
$\Delta y_t = \alpha + \lambda t + \gamma y_{t-1} + v_t$	−3.96	−3.41	−3.13
Standard normal critical values	−2.33	−1.65	−1.28

Note: These critical values are taken from R. Davidson and J. G. MacKinnon, *Estimation and Inference in Econometrics*, New York: Oxford University Press, 1993, p. 708.

Rejection of the null hypothesis that $\gamma = 0$ implies the series is stationary. A failure to reject H_0 suggests the series could be nonstationary, and we must be careful not to proceed to estimate a spurious regression.

To test the hypothesis in (12.22), we estimate the test equation (12.21) by OLS and examine the t-statistic for the hypothesis that $\gamma = 0$. Unfortunately, this t-statistic no longer has the t-distribution that we have used previously to test zero null hypotheses for regression coefficients. The problem arises because, when the null hypothesis is true, y_t is nonstationary and has a variance that increases as the sample size increases. This increasing variance alters the distribution of the usual t-statistic when H_0 is true. To recognize this fact, the statistic is often called a **τ (tau) statistic**, and its value must be compared to specially generated critical values. The critical values are different for each of the variations of the test described in Section 12.3.2. They are tabulated in Table 12.2.[4] Those for test equation (12.21) are given in the middle row. We reject $H_0 : \gamma = 0$ if $\tau \leq \tau_c$, where $\tau = \hat{\gamma}/\text{se}(\hat{\gamma})$ is the OLS "t"-value for $H_0 : \gamma = 0$, and τ_c is a critical value from Table 12.2. In other words, we conclude y_t is stationary if τ is a sufficiently large negative number. Note that the Dickey–Fuller critical values are more negative than the standard normal critical values (shown in the last row). Thus, the τ-statistic must take larger (negative) values than usual for the null hypothesis of nonstationarity ($\gamma = 0$) to be rejected in favor of the alternative of stationarity ($\gamma < 0$).

There are many stationary series that are not adequately modeled by an AR(1) process. A natural question is how do we test for a unit root in a higher-order AR process. It can be shown[5] that testing for a unit root in the AR(p) process

$$y_t = \alpha + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + v_t$$

against the alternative that y_t is stationary, is equivalent to testing $H_0 : \gamma = 0$ against the alternative $H_1 : \gamma < 0$ in the model

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{s=1}^{p-1} a_s \Delta y_{t-s} + v_t \tag{12.23}$$

The original test equation is augmented by the lagged first differences $\Delta y_{t-1} = (y_{t-1} - y_{t-2})$, $\Delta y_{t-2} = (y_{t-2} - y_{t-3})$, … , $\Delta y_{t-p+1} = (y_{t-p+1} - y_{t-p})$. The test procedure for this case uses (12.23) as the test equation but otherwise proceeds just as before, rejecting $H_0 : \gamma = 0$ when $\tau = \hat{\gamma}/\text{se}(\hat{\gamma}) \leq \tau_c$. The critical values are the same as those in Table 12.2. The test is referred to as the **augmented Dickey–Fuller test**. The choice for p can be based on similar criteria to

[4]Originally these critical values were tabulated by the statisticians David Dickey and Wayne Fuller. The values have since been refined, but in deference to the seminal work, unit root tests using these critical values have become known as Dickey–Fuller tests.

[5]See Exercise 12.1.

those described in Chapter 9 for choosing the order of an AR process. Sufficient lags should be included to eliminate autocorrelation in the errors. We can also use significance of the estimates of the a_s, which have their usual large-sample normal distributions, and the AIC and SC variable selection criteria. In practice, we always use the augmented Dickey–Fuller test (rather than the nonaugmented version) to ensure the errors are uncorrelated.

EXAMPLE 12.4 | Checking the Two Interest Rate Series for Stationarity

As an example, consider the two interest rate series—the federal funds rate FFR_t and the three-year bond rate BR_t—plotted in Figures 12.1(e) and (g), respectively. Both series exhibit wandering behavior, wandering up and then down with no discernible trend in either direction. We therefore suspect that they may be nonstationary variables. Using OLS to estimate the test equation (12.23) for each of these variables yields

$$\widehat{\Delta FFR_t} = 0.0580 - 0.0118 FFR_{t-1} + 0.444 \Delta FFR_{t-1} - 0.147 \Delta FFR_{t-2}$$
(τ and t) (−2.47) (12.30) (−4.05)

$$\widehat{\Delta BR_t} = 0.0343 - 0.00635 BR_{t-1} + 0.426 \Delta BR_{t-1} - 0.230 \Delta BR_{t-2}$$
(τ and t) (−1.70) (11.95) (−6.43)

Two augmentation terms have been included for both variables. For *FFR* the number of augmentation terms that minimized the SC was 13—a very large number. However, checking the correlogram of the residuals, we find that including two lags of ΔFFR was sufficient to eliminate any major autocorrelation in the errors. For *BR*, two augmentation terms minimized the SC and were sufficient to eliminate any substantial error autocorrelation. The usual t or normal distributions can be used to assess the significance of the coefficients of the augmentation terms. Their large t-values confirm the decision to include two lags.

However, for checking stationarity, the usual t critical values and p-values cannot be used. Instead, we compare the two τ-values, $\tau = -2.47$ and $\tau = -1.70$ for the coefficients of FFR_{t-1} and BR_{t-1}, respectively, with a critical value from Table 12.2. For a 5% significance level, the relevant critical value is $\tau_{0.05} = -2.86$. The test for stationarity is a one-tail test with the null hypothesis of nonstationarity being rejected if $\tau \leq -2.86$. Since $-2.47 > -2.86$ and $-1.70 > -2.86$, in both cases, we fail to reject H_0. There is insufficient evidence to suggest that *FFR* and *BR* are stationary.

12.3.4 Dickey–Fuller Test with Intercept and Trend

In Sections 12.1.2 and 12.1.3, we introduced two models where a time series y_t has a trend upward or downward. In one, illustrated in Figure 12.4(c), y_t was stationary around a linear trend and described by the process

$$y_t = \alpha + \rho y_{t-1} + \lambda t + v_t \qquad |\rho| < 1 \tag{12.24}$$

A time series that can be described by (12.24) is called trend stationary. The other model was a random walk with drift, illustrated in Figure 12.4(e):

$$y_t = \alpha + y_{t-1} + v_t \tag{12.25}$$

In this case y_t is nonstationary. The Dickey–Fuller test with intercept and trend is designed to discriminate between these two models. Equation (12.25) becomes the null hypothesis (H_0), and equation (12.24) is the alternative hypothesis (H_1). If the null hypothesis is rejected, we conclude y_t is trend stationary. Failure to reject H_0 suggests y_t is nonstationary, or at least there is insufficient evidence to prove otherwise.

Comparing (12.24) and (12.25) suggests a relevant null hypothesis is $H_0: \rho = 1$, $\lambda = 0$. However, like in Section 12.3.3, it has become more common to simply test $H_0: \rho = 1$ against the alternative $H_1: \rho < 1$. A rationale for doing so can be found by going back and checking equation (12.14). There we noted an alternative way of writing (12.24) is

$$(y_t - \mu - \delta t) = \rho(y_{t-1} - \mu - \delta(t-1)) + v_t, \qquad |\rho| < 1$$

where $\mu + \delta t$ is the deterministic trend, $\alpha = \mu(1-\rho) + \rho\delta$ and $\lambda = \delta(1-\rho)$. With these definitions of α and λ, setting $\rho = 1$ implies $\alpha = \delta$ and $\lambda = 0$, giving the random walk with drift in (12.25). As before, the test equation is obtained by subtracting y_{t-1} from both sides of (12.24) and adding augmentation terms to obtain

$$\Delta y_t = \alpha + \gamma y_{t-1} + \lambda t + \sum_{s=1}^{p-1} a_s \Delta y_{t-s} + v_t \tag{12.26}$$

We use the left-tail test $H_0: \gamma = 0$ versus $H_1: \gamma < 0$, rejecting H_0 when $\tau = \hat{\gamma}/\text{se}(\hat{\gamma})$ is less than or equal to a critical value selected from the third row of Table 12.2.

EXAMPLE 12.5 | Is GDP Trend Stationary?

From Figure 12.1(a), we noted that GDP shows a definite upward trend. We now ask whether it can be modeled as stationary around a linear deterministic trend, or whether it contains a stochastic trend component. Using these data to estimate (12.26) yields[6]

$$\widehat{\Delta GDP_t} = \underset{(\tau \text{ and } t)}{0.269} + \underset{}{0.00249t} - \underset{(-2.00)}{0.0330\, GDP_{t-1}} + \underset{(3.58)}{0.312\Delta GDP_{t-1}} + \underset{(2.28)}{0.202\Delta GDP_{t-2}}$$

Two augmentation terms minimized the SC, eliminated major autocorrelation in the residuals, and had coefficient estimates significant at a 5% level. For assessing stationarity, we find $\tau = -2.00$, which is greater than the 5% critical value $\tau_{0.05} = -3.41$. Thus, we cannot reject the null hypothesis that GDP follows a nonstationary random walk with drift. There is insufficient evidence to conclude that GDP is trend stationary.

EXAMPLE 12.6 | Is Wheat Yield Trend Stationary?

In Example 12.2, we model wheat yield in the Toodyay Shire of Western Australia with a deterministic trend. To see whether this choice was justified we estimate the test equation

$$\widehat{\Delta \ln(YIELD_t)} = \underset{(\tau)}{-0.158} + \underset{}{0.0167\, t} - \underset{(-5.24)}{0.745 \ln(YIELD_{t-1})}$$

In this case, no augmentation terms were necessary. The value $\tau = -5.24$ is less than the 5% critical value $\tau_{0.05} = -3.41$ and so, at this level of significance, we reject a null hypothesis of nonstationarity and conclude that $\ln(YIELD)$ is trend stationary.

12.3.5 Dickey–Fuller Test with No Intercept and No Trend

In its simplest form with no augmentation terms, this test is designed to test the null hypothesis of a random walk $H_0: y_t = y_{t-1} + v_t$ against the stationary AR(1) alternative $H_1: y_t = \rho y_{t-1} + v_t$, $|\rho| < 1$. Since y_t has a zero mean when H_1 is true, it is designed for series that are centered around zero, like that in Figure 12.4(a). The test equation is

$$\Delta y_t = \gamma y_{t-1} + \sum_{s=1}^{p-1} a_s \Delta y_{t-s} + v_t \tag{12.27}$$

[6]The trend term takes the values 0, 1, 2, …,132 with 1984Q1 = 0.

TABLE 12.3 **AR Processes and the Dickey–Fuller Tests**

AR Processes: $\|\rho\| < 1$	Setting $\rho = 1$	Dickey–Fuller Tests
$y_t = \rho y_{t-1} + u_t$	$y_t = y_{t-1} + u_t$	Test with no constant and no trend
$y_t = \alpha + \rho y_{t-1} + v_t$ $\alpha = \mu(1-\rho)$	$y_t = y_{t-1} + v_t$ $\alpha = 0$	Test with constant and no trend
$y_t = \alpha + \rho y_{t-1} + \lambda t + v_t$ $\alpha = \mu(1-\rho) + \rho\delta$ $\lambda = \delta(1-\rho)$	$y_t = \delta + y_{t-1} + v_t$ $\alpha = \delta$ $\lambda = 0$	Test with constant and trend

We test $H_0: \gamma = 0$ against $H_1: \gamma < 0$ as described previously, and the critical values are given in the first row of Table 12.2.

Most time series measured in terms of their original levels do not have a zero mean. However, their first differences $\Delta y_t = y_t - y_{t-1}$ may turn out to have a zero mean. For example, the first difference of the random walk $y_t = y_{t-1} + v_t$ is $\Delta y_t = v_t$ which has a zero mean. Testing whether first differences are stationary has relevance for finding the **order of integration** of a series which we consider in Section 12.3.6.

In Table 12.3, we summarize the models under H_0 and H_1 for each of the three tests, omitting the augmentation terms to avoid cluttering the table.

12.3.6 Order of Integration

Up to this stage, we have discussed only whether a series is stationary or nonstationary. We can take the analysis another step forward and consider a concept called the "order of integration." Recall that if y_t follows a random walk, then $\gamma = 0$ and the first difference of y_t becomes

$$\Delta y_t = y_t - y_{t-1} = v_t$$

An interesting feature of the series $\Delta y_t = y_t - y_{t-1}$ is that it is stationary since v_t, being an independent $(0, \sigma_v^2)$ random variable, is stationary. Series like y_t, which can be made stationary by taking the first difference, are said to be **integrated of order one**, and denoted as **I(1)**. Stationary series are said to be integrated of order zero, **I(0)**. In general, the order of integration of a series is the minimum number of times it must be differenced to make it stationary.

EXAMPLE 12.7 | The Order of Integration of the Two Interest Rate Series

In Example 12.4, we concluded that the two interest rate series *FFR* and *BR* were nonstationary. To find their order of integration, we ask the next question: are their first differences, $\Delta FFR_t = FFR_t - FFR_{t-1}$ and $\Delta BR_t = BR_t - BR_{t-1}$ stationary? Their plots, in Figures 12.1(f) and (h), suggest stationarity. Given these plots appear to fluctuate around zero, we use the Dickey–Fuller test equation with no intercept and no trend, to obtain the following results.

$$\widehat{\Delta(\Delta FFR_t)} = -0.715\Delta FFR_{t-1} + 0.157\Delta(\Delta FFR_{t-1})$$
$$(\tau \text{ and } t) \quad (-17.76) \qquad (4.33)$$

$$\widehat{\Delta(\Delta BR_t)} = -0.811\Delta BR_{t-1} + 0.235\Delta(\Delta BR_{t-1})$$
$$(\tau \text{ and } t) \quad (-19.84) \qquad (6.58)$$

where $\Delta(\Delta FFR_t) = \Delta FFR_t - \Delta FFR_{t-1}$ and $\Delta(\Delta BR_t) = \Delta BR_t - \Delta BR_{t-1}$. In both cases, one augmentation term was sufficient to eliminate serial correlation in the errors. Note that the null hypotheses are that the variables ΔF and ΔB are not stationary. The large negative values of the τ-statistic, $\tau = -17.76$ for ΔFFR and $\tau = -19.84$ for ΔBR, are much

less and the 5% critical value $\tau_{0.05} = -1.94$. We therefore reject null hypotheses that ΔFFR and ΔBR have unit roots and conclude they are stationary.

These results imply that, while the levels of the two interest rates are nonstationary, their first differences are stationary. We say that the series FFR_t and BR_t are I(1) because they had to be differenced once to make them stationary $\left[\Delta FFR_t \text{ and } \Delta BR_t \text{ are I(0)}\right]$. In the Sections 12.4 and 12.5, we investigate the implications of these results for regression modeling.

12.3.7 Other Unit Root Tests

While augmented Dickey–Fuller tests remain the most popular tests for unit roots, the power of the tests is low in the sense that they often cannot distinguish between a highly persistent stationary process (where ρ is very close but not equal to 1) and a nonstationary process (where $\rho = 1$). The power of the test also diminishes as deterministic terms constant and trend are included in the test equation. Here we briefly mention other tests that have been developed with a view to improving the power of the test: the Elliot, Rothenberg, and Stock (ERS), Phillips and Perron (PP), Kwiatkowski, Phillips, Schmidt, and Shin (KPSS), and Ng and Perron (NP) tests.[7] Each test carries an abbreviation from the names of its developers.

The ERS test proposes removing the constant/trend effects from the data and performing the unit root test on the residuals. The distribution of the t-statistic is now devoid of deterministic terms (i.e., the constant and/or trend). The PP test adopts a nonparametric approach that assumes a general autoregressive moving-average structure and uses spectral methods to estimate the standard error of the test correlation. Instead of specifying a null hypothesis of nonstationary, the KPSS test specifies a null hypothesis that the series is stationary or trend stationary. NP tests suggest various modifications of the PP and ERS tests.

12.4 Cointegration

As a general rule, to avoid the problem of spurious regression, nonstationary time-series variables should not be used in regression models. However, there is an exception to this rule. If y_t and x_t are nonstationary I(1) variables, then we expect their difference, or any linear combination of them, such as $e_t = y_t - \beta_1 - \beta_2 x_t$,[8] to be I(1) as well. However, there is an important case when $e_t = y_t - \beta_1 - \beta_2 x_t$ is a stationary I(0) process. In this case, y_t and x_t are said to be **cointegrated**. Cointegration implies that y_t and x_t share similar stochastic trends, and, since the difference e_t is stationary, they never diverge too far from each other.

A natural way to test whether y_t and x_t are cointegrated is to test whether the errors $e_t = y_t - \beta_1 - \beta_2 x_t$ are stationary. Since we cannot observe e_t, we test the stationarity of the OLS residuals, $\hat{e}_t = y_t - b_1 - b_2 x_t$ using a Dickey–Fuller test. The test for cointegration is effectively a test of the stationarity of the residuals. If the residuals are stationary, then y_t and x_t are said to be cointegrated; if the residuals are nonstationary, then y_t and x_t are not cointegrated, and any apparent regression relationship between them is said to be spurious.

The test for stationarity of the residuals is based on the test equation

$$\Delta\hat{e}_t = \gamma\hat{e}_{t-1} + v_t \tag{12.28}$$

where $\Delta\hat{e}_t = \hat{e}_t - \hat{e}_{t-1}$. As before, we examine the t (or *tau*) statistic for the estimated slope coefficient. Note that the regression has no constant term because the mean of the regression residuals

[7] More details can be found in William Greene, *Econometric Analysis*, 8th ed., Chapter 21, 2018, Pearson.

[8] A linear combination of x and y is a new variable $z = a_0 + a_1 x + a_2 y$. Here we set the constants $a_0 = -\beta_1$, $a_1 = -\beta_2$, and $a_2 = 1$ and call z the series e.

TABLE 12.4 **Critical Values for the Cointegration Test**

Regression Model	1%	5%	10%
(1) $y_t = \beta x_t + e_t$	−3.39	−2.76	−2.45
(2) $y_t = \beta_1 + \beta_2 x_t + e_t$	−3.96	−3.37	−3.07
(3) $y_t = \beta_1 + \delta t + \beta_2 x_t + e_t$	−3.98	−3.42	−3.13

Note: These critical values are taken from J. Hamilton, *Time Series Analysis*, Princeton University Press, 1994, p. 766.

is zero. Also, since we are basing this test upon **estimated** values of the residuals, the critical values will be different from those in Table 12.2. The proper critical values for a test of cointegration are given in Table 12.4. The test equation can also include extra terms like $\Delta\hat{e}_{t-1}, \Delta\hat{e}_{t-2}, \ldots$ on the right-hand side if they are needed to eliminate autocorrelation in v_t.

There are three sets of critical values. Which set we use depends on whether the residuals $\hat{e}_t$ are derived from a regression equation without a constant term [like (12.29a)] or a regression equation with a constant term [like (12.29b)], or a regression equation with a constant and a time trend [like (12.29c)].

$$\text{Equation 1}: \ \hat{e}_t = y_t - bx_t \tag{12.29a}$$

$$\text{Equation 2}: \ \hat{e}_t = y_t - b_2 x_t - b_1 \tag{12.29b}$$

$$\text{Equation 3}: \ \hat{e}_t = y_t - b_2 x_t - b_1 - \hat{\delta} t \tag{12.29c}$$

EXAMPLE 12.8 | Are the Federal Funds Rate and Bond Rate Cointegrated?

To illustrate, let us test whether $y_t = BR_t$ and $x_t = FFR_t$, as plotted in Figures 12.1(e) and (g), are cointegrated. We have already shown that both series are nonstationary. The estimated least-squares regression between these variables is

$$\widehat{BR}_t = 1.328 + 0.832\, FFR_t \qquad R^2 = 0.908$$
$$(t) \qquad (85.72) \qquad\qquad (12.30)$$

The estimated test equation for stationarity in the OLS residuals $\hat{e}_t = BR_t - 1.328 - 0.832\, FFR_t$ is

$$\widehat{\Delta\hat{e}_t} = -0.0817\,\hat{e}_{t-1} + 0.223\,\Delta\hat{e}_{t-1} - 0.177\,\Delta\hat{e}_{t-2}$$
$$(\tau \text{ and } t) \quad (-5.53) \qquad (6.29) \qquad (-4.90)$$

Note that this is the augmented Dickey–Fuller version of the test with two lagged terms Δe_{t-1} and Δe_{t-2} to correct for autocorrelation. Since there is a constant term in (12.30), we use the equation (2) critical values in Table 12.4.

The null and alternative hypotheses in the test for cointegration are

H_0 : the series are not cointegrated
$\iff$ residuals are nonstationary

H_1 : the series are cointegrated
$\iff$ residuals are stationary

Similar to the one-tail unit root tests, we reject the null hypothesis of no cointegration if $\tau \le \tau_c$, and we do not reject the null hypothesis that the series are not cointegrated if $\tau > \tau_c$. The tau statistic in this case is −5.53 which is less than the critical value −3.37 at the 5% level of significance. Thus, we reject the null hypothesis that the least-squares residuals are nonstationary and conclude that they are stationary. This implies that the bond rate and the federal funds rate are cointegrated. In other words, there is a fundamental relationship between these two variables (the estimated regression relationship between them is valid and not spurious) and the estimated values of the intercept and slope are 1.328 and 0.832, respectively.

The result—that the federal funds and bond rates are cointegrated—has major economic implications! It means that when the Federal Reserve implements monetary policy by changing the federal funds rate, the bond rate will also change thereby ensuring that the effects of monetary policy are transmitted to the rest of the economy. In contrast, the effectiveness of monetary policy would be severely hampered if the bond and federal funds rates were spuriously related as this implies that their movements, fundamentally, have little to do with each other.

12.4.1 The Error Correction Model

In Section 12.4, we discussed the concept of cointegration as the relationship between I(1) variables such that the residuals are I(0). A relationship between I(1) variables is also often referred to as a long-run relationship while a relationship between I(0) variables is often referred to as a short-run relationship. In this section, we describe a dynamic relationship between I(0) variables, which embeds a cointegrating relationship, known as the short-run error correction model.

As discussed in Chapter 9, when one is working with time-series data, it is quite common, and in fact, quite important to allow for dynamic effects. To derive the error correction model requires a bit of algebra, but we shall persevere as this model offers a coherent way to combine the long- and short-run effects.

Let us start with a general model that contains lags of y and x, namely the ARDL model introduced in Chapter 9, except that now the variables are nonstationary:

$$y_t = \delta + \theta_1 y_{t-1} + \delta_0 x_t + \delta_1 x_{t-1} + v_t$$

For simplicity, we shall consider lags up to order one, but the following analysis holds for any order of lags. Now recognize that if y and x are cointegrated, it means that there is a long-run relationship between them. To derive this exact relationship, we set $y_t = y_{t-1} = y$, $x_t = x_{t-1} = x$ and $v_t = 0$ and then, imposing this concept in the ARDL, we obtain

$$y(1-\theta_1) = \delta + (\delta_0 + \delta_1)x$$

This equation can be rewritten as $y = \beta_1 + \beta_2 x$ where $\beta_1 = \delta/(1-\theta_1)$ and $\beta_2 = (\delta_0 + \delta_1)/(1-\theta_1)$. To repeat, we have now derived the implied cointegrating relationship between y and x; alternatively, we have derived the long-run relationship that holds between the two I(1) variables.

We will now manipulate the ARDL to see how it embeds the cointegrating relation. First, add the term $-y_{t-1}$ to both sides of the equation:

$$y_t - y_{t-1} = \delta + (\theta_1 - 1)y_{t-1} + \delta_0 x_t + \delta_1 x_{t-1} + v_t$$

Second, add the term $-\delta_0 x_{t-1} + \delta_0 x_{t-1}$ to the right-hand side to obtain

$$\Delta y_t = \delta + (\theta_1 - 1)y_{t-1} + \delta_0(x_t - x_{t-1}) + (\delta_0 + \delta_1)x_{t-1} + v_t$$

where $\Delta y_t = y_t - y_{t-1}$. If we then manipulate the equation to look like

$$\Delta y_t = (\theta_1 - 1)\left(\frac{\delta}{(\theta_1 - 1)} + y_{t-1} + \frac{(\delta_0 + \delta_1)}{(\theta_1 - 1)}x_{t-1}\right) + \delta_0 \Delta x_t + v_t$$

where $\Delta x_t = x_t - x_{t-1}$, and do a bit more tidying, using the definitions β_1 and β_2, we get

$$\Delta y_t = -\alpha(y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + \delta_0 \Delta x_t + v_t \tag{12.31}$$

where $\alpha = (1-\theta_1)$. As you can see, the expression in parenthesis is the cointegrating relationship. In other words, we have embedded the cointegrating relationship between y and x in a general ARDL framework.

Equation (12.31) is called an error correction equation because (a) the expression $(y_{t-1} - \beta_1 - \beta_2 x_{t-1})$ shows the deviation of y_{t-1} from its long-run value, $\beta_1 + \beta_2 x_{t-1}$—in other words, the "error" in the previous period—and (b) the term $(\theta_1 - 1)$ shows the "correction" of Δy_t to the "error." More specifically, if the error in the previous period is positive so that $y_{t-1} > (\beta_1 + \beta_2 x_{t-1})$, then y_t should fall and Δy_t should be negative; conversely, if the error in the previous period is negative so that $y_{t-1} < (\beta_1 + \beta_2 x_{t-1})$, then y_t should rise and Δy_t should be positive. This means that if a cointegrating relationship between y and x exists, so that adjustments always work to "error-correct," then empirically we should also find that $(1-\theta_1) > 0$, which implies that $\theta_1 < 1$. If there is no evidence of cointegration between the variables, then the estimate for θ_1 would be insignificant.

The error correction model is a very popular model because it allows for the existence of an underlying or fundamental link between variables (the long-run relationship) as well as for short-run adjustments (i.e., changes) between variables, including adjustments toward the cointegrating relationship. It also shows that we can work with I(1) variables (y_{t-1}, x_{t-1}) and I(0) variables $(\Delta y_t, \Delta x_t)$ in the same equation provided that (y, x) are cointegrated, meaning that the term $(y_{t-1} - \beta_0 - \beta_1 x_{t-1})$ contains stationary residuals. In fact, this formulation can also be used to test for cointegration between y and x.

To estimate (12.31) we can proceed in one of two ways: we can estimate the equation with $y_{t-1} - \beta_1 - \beta_2 x_{t-1}$ replaced by $\hat{e}_{t-1}$, or we can find new estimates of β_1 and β_2 at the same time as we estimate α and δ_0. For the latter approach, we can estimate the parameters directly by applying nonlinear least squares to (12.31), or we can use OLS to estimate the equation

$$\Delta y_t = \beta_1^* + \alpha^* y_{t-1} + \beta_2^* x_{t-1} + \delta_0 \Delta x_{t-1} + v_t$$

and retrieve the parameters in equation (12.31) from $\alpha = -\alpha^*$, $\beta_1 = -\beta_1^*/\alpha^*$ and $\beta_2 = -\beta_2^*/\alpha^*$. The nonlinear least squares and the retrieved OLS estimates will be identical. However, they will differ slightly from the two-step estimates obtained by replacing $y_{t-1} - \beta_1 - \beta_2 x_{t-1}$ with $\hat{e}_{t-1}$.

EXAMPLE 12.9 | An Error Correction Model for the Bond and Federal Funds Rates

For an error correction model relating changes in the bond rate to the lagged cointegrating relationship and changes in the federal funds rate, it turns out that up to four lags of ΔFFR_t are relevant and two lags of ΔBR_t are needed to eliminate serial correlation in the error. The equation estimated directly using nonlinear least squares is

$$\begin{aligned}
\widehat{\Delta BR_t} &= -0.0464\left(BR_{t-1} - 1.323 - 0.833 FFR_{t-1}\right) \\
(t) \qquad & (3.90) \\
&+ 0.272\Delta BR_{t-1} - 0.242\Delta BR_{t-2} \\
& (7.27) \qquad (-6.40) \\
&+ 0.342\Delta FFR_t - 0.105\Delta FFR_{t-1} + 0.099\Delta FFR_{t-2} \\
& (14.22) \qquad (-3.83) \qquad (3.62) \\
&- 0.066\Delta FFR_{t-3} + 0.056\Delta FFR_{t-4} \\
& (-2.69) \qquad (2.46) \qquad\qquad (12.32)
\end{aligned}$$

Notice that the estimates $\hat{\beta}_1 = 1.323$ and $\hat{\beta}_2 = 0.833$ are very similar to those obtained from direct OLS estimation of the cointegrating relationship in (12.30). The relationship between all the coefficients in (12.32) and its corresponding ARDL model are explored in Exercise 12.18.

If we use the residuals $\hat{e}_t = BR_t - 1.323 - 0.833 FFR_t$, obtained from the estimates in (12.32), to test for cointegration, we get a similar result to our earlier one

$$\begin{aligned}
\widehat{\Delta e_t} &= -0.0819\hat{e}_{t-1} + 0.224\Delta\hat{e}_{t-1} - 0.177\Delta\hat{e}_{t-2} \\
(\tau \text{ and } t) \quad & (-5.53) \qquad (6.29) \qquad (-4.90)
\end{aligned}$$

As before, the null hypothesis is that (BR, FFR) are not cointegrated (the residuals are nonstationary). Since the cointegrating relationship includes a constant, the critical value from Table 12.4 is -3.37. Comparing the actual value $\tau = -5.53$ with the critical value, we reject the null hypothesis and conclude that (BR, FFR) are cointegrated.

12.5 Regression When There Is No Cointegration

Thus far, we have shown that regression with I(1) variables is acceptable providing those variables are cointegrated, allowing us to avoid the problem of spurious results. We also know that regression with stationary I(0) variables, that we studied in Chapter 9, is acceptable. What happens when there is no cointegration between I(1) variables? In this case, the sensible thing to do is to convert the nonstationary series to stationary series and to use the techniques discussed in Chapter 9 to estimate dynamic relationships between the stationary variables. However, we stress that this step should be taken only when we fail to find cointegration between the I(1) variables.

Regression with cointegrated I(1) variables makes the least-squares estimator "super-consistent"[9] and, moreover, it is economically useful to establish relationships between the levels of economic variables.

How we convert nonstationary series to stationary series, and the kind of model we estimate, depend on whether the variables are **difference stationary** or **trend stationary**. In the former case, we convert the nonstationary series to its stationary counterpart by taking first differences. We dealt with the latter case in Section 12.1.1 where we converted the nonstationary series to a stationary series by detrending, or we included a trend term in the regression relationship. We now consider how to estimate regression relationships with nonstationary variables that are neither cointegrated nor trend stationary.

Recall that an I(1) variable is one that is stationary after differencing once. Another name for variables with this characteristic is that they are **first-difference stationary**. Specifically, if y_t is nonstationary with a stochastic trend and its first difference $\Delta y_t = y_t - y_{t-1}$ is stationary, then y_t is I(1) and first-difference stationary. If Dickey–Fuller tests reveal that two variables, y and x, that you would like to relate in a regression, are first difference stationary and not cointegrated, then a suitable regression involving only stationary variables is one that relates changes in y to changes in x, with relevant lags included. If y_t and x_t behave like random walks with no obvious trend, then an intercept can be omitted. For example, using one lagged Δy_t and a current and lagged Δx_t, we have:

$$\Delta y_t = \theta \Delta y_{t-1} + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + e_t \tag{12.33}$$

If y_t and x_t behave like random walks with drift, then it is appropriate to include an intercept, an example of which is

$$\Delta y_t = \alpha + \theta \Delta y_{t-1} + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + e_t \tag{12.34}$$

Note that a random walk with drift is such that $\Delta y_t = \alpha + v_t$, implying an intercept should be included, whereas a random walk with no drift becomes $\Delta y_t = v_t$. In line with Chapter 9, the models in (12.33) and (12.34) are ARDL models with first-differenced variables. In general, since there is often doubt about the role of the constant term, the usual practice is to include an intercept term in the regression.

EXAMPLE 12.10 | A Consumption Function in First Differences

In Chapter 9, there were a number of examples and exercises involving first differences of variables. When studying that chapter, you may have wondered why we did not use variables in their levels. The reason is now clear. It was to ensure the variables were stationary. In the following example of a consumption function, we return to the data file *cons_inc*, containing quarterly data on Australian consumption expenditure and national disposable income, used earlier in Example 9.16. We will use data from 1985Q1 to 2016Q3. Plots of the series appear in Figure 12.6.

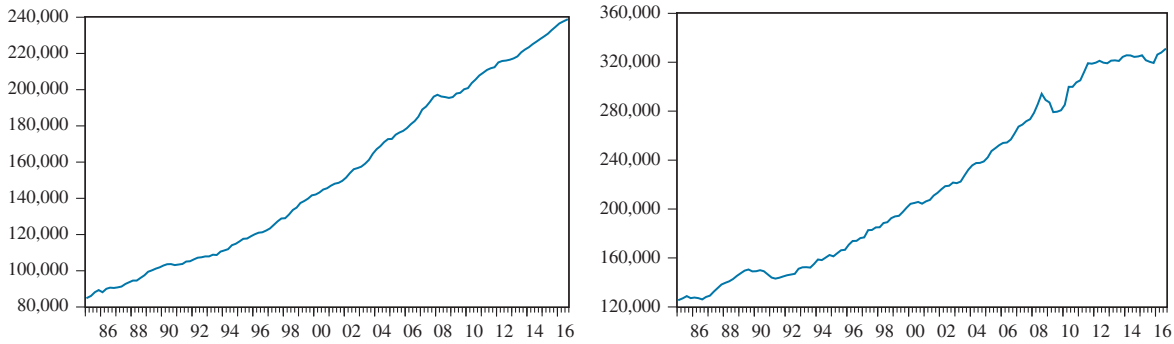
Since both consumption (C) and income (Y) are clearly trending, we include a trend term in the Dickey–Fuller test equations to see if they should be treated as trend stationary or difference stationary. The results from the test equations are

$$\widehat{\Delta C_t} = 1989.7 + 29.43\,t - 0.0193 C_{t-1} + 0.244 \Delta C_{t-1}$$
$$(\tau \text{ and } t) \qquad (2.03)\ (-1.70) \qquad (2.82)$$

$$\widehat{\Delta Y_t} = 5044.6 + 80.04\,t - 0.0409 Y_{t-1} + 0.248 \Delta Y_{t-1}$$
$$(\tau \text{ and } t) \qquad (2.27)\ (-2.14) \qquad (2.89)$$

From Table 12.2, the 5% critical value for test equations that include a trend is $\tau_{0.05} = -3.41$. The τ values for consumption (-1.70) and income (-2.14) are both greater than $\tau_{0.05}$. Hence, we are unable to conclude that C and Y are trend stationary.

[9] Consistency means that as $T \to \infty$ the least squares estimator converges to the true parameter value. See Section 5.7. Super-consistency means that it converges to the true value at a faster rate.



(*a*) Consumption (C) (*b*) Income (Y)

FIGURE 12.6 **Australian consumption and disposable income.**

The next step is to see if C and Y are cointegrated. Because they are trending, we include a trend term, and estimate the following equation, saving the residuals.

$$\hat{C}_t = -18746 + 420.4t + 0.468Y_t$$
$$(t) \qquad (9.92) \quad (20.49) \qquad (12.35)$$

If the residuals are stationary, we conclude C and Y are cointegrated and (12.35) is a valid regression. If the residuals are nonstationary, then (12.35) could be a spurious regression. The test equation for assessing the stationarity of the residuals is

$$\widehat{\Delta\hat{e}_t} = -0.121\hat{e}_{t-1} + 0.263\Delta\hat{e}_{t-1}$$
$$(\tau \text{ and } t) \ (-2.93) \qquad (2.94)$$

Comparing $\tau = -2.93$ with the critical value of $\tau_{0.05} = -3.42$ in the third row of Table 12.4, we fail to reject a null hypothesis that the residuals are nonstationary (C and Y are not cointegrated).

Having established that C and Y are not trend stationary and not cointegrated, or at least that there is insufficient evidence to suggest otherwise, the natural regression to estimate relating the two variables is one in first differences. First, however, we need to confirm that they are first-difference stationary (integrated of order one). The unit-root test equations for this purpose are

$$\widehat{\Delta(\Delta C_t)} = 844.0 - 0.689\Delta C_{t-1}$$
$$(\tau) \qquad (-8.14)$$

$$\widehat{\Delta(\Delta Y_t)} = 1{,}228.7 - 0.751\Delta Y_{t-1}$$
$$(\tau) \qquad (-8.68)$$

We include a constant in these equations because the unit-root test for the variables in their levels included a trend. The test values $\tau = -8.14$ and $\tau = -8.68$ are less than 5% critical value $\tau_{0.05} = -2.86$ from Table 12.2. We therefore conclude that ΔC and ΔY are stationary and hence that C and Y are first-difference stationary. Proceeding to estimate an ARDL model for C and Y in first differences, we obtain

$$\widehat{\Delta C_t} = 785.8 + 0.0573\Delta Y_t + 0.282\Delta C_{t-1}$$
$$(t) \qquad (2.07) \qquad (3.34)$$

12.6 Summary

- If variables are stationary, or I(1) and cointegrated, we can estimate a regression relationship between the levels of those variables without fear of encountering a spurious regression. In the latter case, we can do this by estimating a least-squares equation between the I(1) variables or by estimating a nonlinear least-squares error correction model which embeds the I(1) variables.
- If the variables are I(1) and not cointegrated, we need to estimate a relationship in first differences, with or without the constant term.

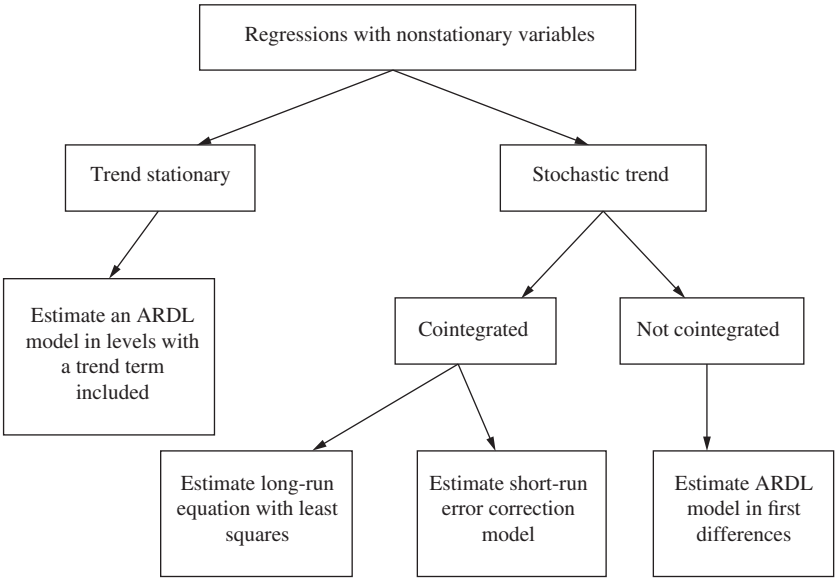


FIGURE 12.7 **Regression with time-series data: nonstationary variables.**

- If they are trend stationary, we can either detrend the series first and then perform regression analysis with the stationary (detrended) variables or, alternatively, estimate a regression relationship that includes a trend variable.

These options are shown in Figure 12.7.

12.7 Exercises

12.7.1 Problems

12.1 Consider the AR(2) model $y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + v_t$. Suppose that

$$1 - \theta_1 z - \theta_2 z^2 = (1 - c_1 z)(1 - c_2 z)$$

a. Show that $c_1 + c_2 = \theta_1$ and $c_1 c_2 = -\theta_2$.

b. Prove that the AR(2) model has a unit root if and only if $\theta_1 + \theta_2 - 1 = 0$. [*Hint*: The roots of $1 - \theta_1 z - \theta_2 z^2 = 0$ are $1/c_1$ and $1/c_2$.]

c. Prove that $\theta_1 + \theta_2 - 1 < 0$ if the AR(2) process is stationary.

d. Prove that the AR(2) model $y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + v_t$ can also be written as

$$\Delta y_t = \delta + \gamma y_{t-1} + a_1 \Delta y_{t-1} + v_t$$

where $\gamma = \theta_1 + \theta_2 - 1$ and $a_1 = -\theta_2$. What are the implications of this result and the results in parts (b) and (c) for unit root tests in an AR(2) model.

e. Show that an AR(p) model has a unit root if $\gamma = \theta_1 + \theta_2 + \cdots + \theta_p - 1 = 0$.

f. Show that setting $\gamma = \theta_1 + \theta_2 + \cdots + \theta_p - 1$ in equation (12.23) implies $a_j = -\sum_{r=j}^{p-1} \theta_{r+1}$.

12.2 **a.** Consider the stationary AR(1) model $y_t = \rho y_{t-1} + v_t$, $|\rho| < 1$. The v_t are independent random errors with mean zero and variance σ_v^2. In Appendix 9B we showed that the autocorrelations for this model are given by $\text{corr}(y_t, y_{t+s}) = \rho^s$. Given $\rho = 0.9$, find the autocorrelations for observations 1 period apart, 2 periods apart, etc., up to 10 periods apart.

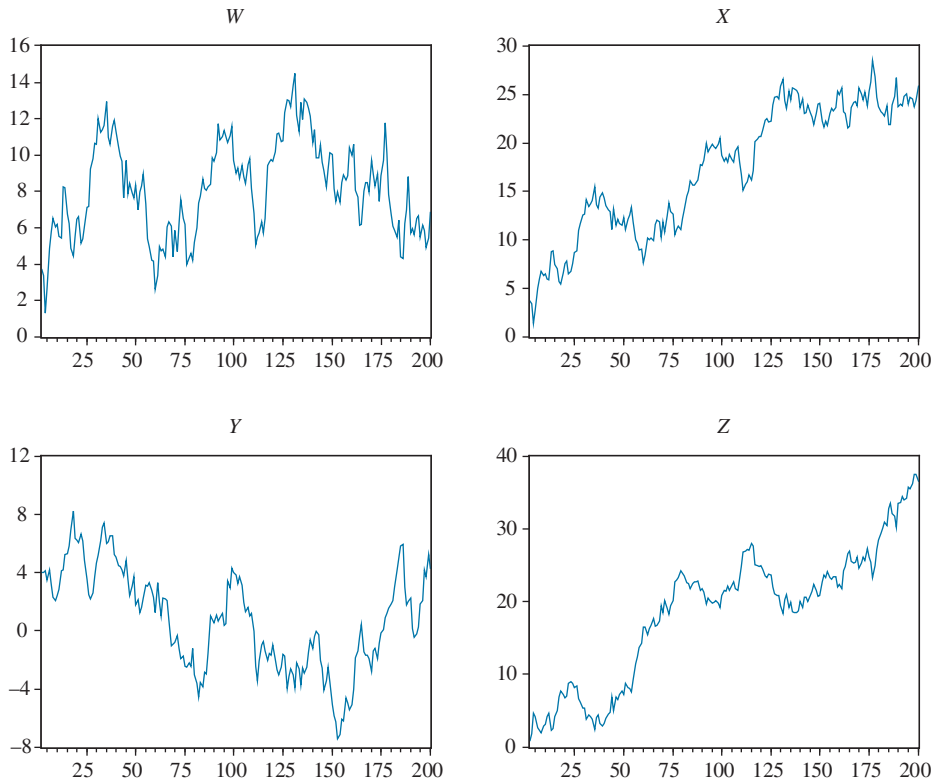


FIGURE 12.8 **Time series for Exercise 12.3.**

b. Consider the nonstationary random walk model $y_t = y_{t-1} + v_t$. Assuming a fixed $y_0 = 0$, rewrite y_t as a function of all past errors $v_{t-1}, v_{t-2}, \ldots, v_1$.

c. Use the result in part (b) to find (i) the mean of y_t, (ii) the variance of y_t, and (iii) the covariance between y_t and y_{t+s}.

d. Use the results from part (c) to show that $\text{corr}(y_t, y_{t+s}) = \sqrt{t/(t+s)}$.

e. Assume $t = 100$ (the random walk has been operating for 100 periods). Find the correlations between y_{100} and y in each of the next 10 periods (up to y_{110}). Compare these correlations with those obtained in part (a).

f. Find corr(y_{100}, y_{200}) for each of the two models and comment on their magnitudes.

12.3 Figure 12.8 shows plots of four time series that are stored in the data file *unit*.

a. The results from Dickey–Fuller test equations for these four variables are given below. Explain why these equations were chosen. No augmentation terms are included. What criteria would have led to their omission?

$$\widehat{\Delta W_t} = 0.778 - 0.0936 W_{t-1}$$
$$(\tau) \qquad (-3.23)$$
$$\widehat{\Delta Y_t} = 0.0304 - 0.0396 Y_{t-1}$$
$$(\tau) \qquad (-1.98)$$
$$\widehat{\Delta X_t} = 0.805 - 0.0939 X_{t-1} + 0.00928\, t$$
$$(\tau) \qquad (-3.13)$$
$$\widehat{\Delta Z_t} = 0.318 - 0.0355 Z_{t-1} + 0.00306\, t$$
$$(\tau) \qquad (-1.87)$$

b. Carry out Dickey–Fuller tests on each of the series. What do you conclude about their stationarity properties?

c. The following estimated equation uses residuals $\hat{e}_t$ from least-squares estimation of the model $X_t = \beta_1 + \delta t + \beta_2 Z_t + e_t$. Can you conclude that X_t and Z_t are cointegrated?

$$\widehat{\Delta \hat{e}_t} = -0.0683 \hat{e}_{t-1}$$
$$(\tau) \quad (-2.83)$$

d. Use the following equation and earlier results to assess the order of integration of Z_t.

$$\widehat{\Delta(\Delta Z_t)} = 0.174 - 0.987 \Delta Z_{t-1}$$
$$(\tau) \quad (-13.76)$$

12.4 A time-series process of the form $y_t = \alpha + y_{t-1} + v_t$, $v_t \sim N(0, \sigma^2)$ can be rearranged as $y_t - y_{t-1} = \Delta y_t = \alpha + v_t$. This shows that y_t is integrated of order one, since its first difference is stationary. Show that a time series of the form $y_t = 2y_{t-1} - y_{t-2} + \alpha + v_t$ is integrated of order two.

12.5 In Chapter 9, we found that, given a time series of observations, $I_T = \{(y_1, x_1), (y_2, x_2), \ldots, (y_T, x_T)\}$, the best one-period and two-period ahead forecasts for y_{T+1} and y_{T+2} were given by $E(y_{T+1}|I_T)$ and $E(y_{T+2}|I_T)$, respectively. Given that $T = 29$, $y_T = 10$, $y_{T-1} = 12$, $x_{T+2} = x_{T+1} = x_T = 5$, and $x_{T-1} = 6$, find forecasts for y_{T+1} and y_{T+2} from each of the following models. In each case, assume that v_t are independent random errors distributed as $N(0, \sigma_v^2 = 4)$.

a. The random walk $y_t = y_{t-1} + v_t$.

b. The random walk with drift $y_t = 5 + y_{t-1} + v_t$.

c. The random walk $\ln(y_t) = \ln(y_{t-1}) + v_t$.

d. The deterministic trend model $y_t = 10 + 0.1\,t + v_t$.

e. The ARDL model $y_t = 6 + 0.6y_{t-1} + 0.3x_t + 0.1x_{t-1} + v_t$.

f. The error correction model $\Delta y_t = -0.4(y_{t-1} - 15 - x_{t-1}) + 0.3\Delta x_t + v_t$. In addition, find the long-run equilibrium value for y when $x = 5$.

g. The first difference model $\Delta y_t = 0.6\Delta y_{t-1} + 0.3\Delta x_t + 0.1\Delta x_{t-1} + v_t$.

12.6 Increases in the mortgage interest rate increase the cost of owning a house and lower the demand for houses. In this question we investigate the properties of two time series that could be used to model this demand relationship: the number of new one-family houses sold in the U.S. (*HOMES*) and the 30-year conventional mortgage rate (*IRATE*). These series, along with their changes, *DHOMES* and *DIRATE*, are plotted in Figure 12.9. The data are from January 1992 (1992M1) to September 2016 (2016M9). The units of measurement are thousands of new houses for *HOMES* and percentage points for *IRATE*.

Use the following test equation results to test for unit roots. In each case give the null and alternative hypotheses, and draw a conclusion. In all cases use a 5% significance level. Based on the test results, describe how you would set up a model for the demand relationship. (In each model, the OLS standard errors do not reflect the true variance of the estimator $\hat{\gamma}$ but, nevertheless, they can be used to construct the τ-statistic.)

a.
$$\widehat{\Delta HOMES_t} = 7.051 - 0.0102 HOMES_{t-1} - 0.280 \Delta HOMES_{t-1}$$
$$(\text{se}) \quad (0.0096) \quad (0.056)$$

b.
$$\widehat{\Delta HOMES_t} = 16.36 - 0.0385\,t - 0.0151 HOMES_{t-1} - 0.279 \Delta HOMES_{t-1}$$
$$(\text{se}) \quad (0.0345) \quad (0.0106) \quad (0.056)$$

c.
$$\widehat{\Delta IRATE_t} = 0.0477 - 0.00985 IRATE_{t-1} + 0.300 \Delta IRATE_{t-1}$$
$$(\text{se}) \quad (0.00679) \quad (0.056)$$

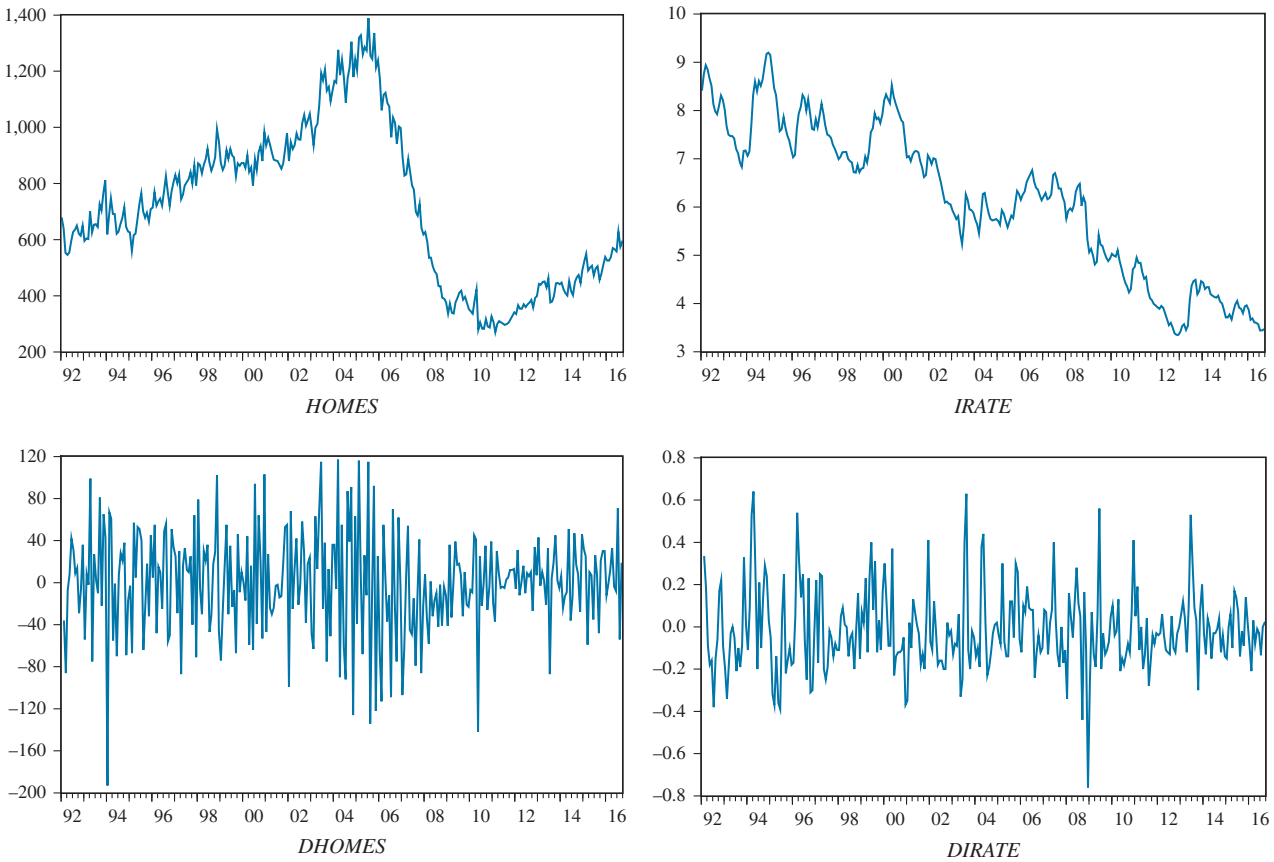


FIGURE 12.9 **Time series for new houses and the mortgage rate and their changes.**

d. $\widehat{\Delta IRATE_t} = 0.603 - 0.00120t - 0.0710IRATE_{t-1} + 0.329\Delta IRATE_{t-1}$
(se) (0.00033) (0.0181) (0.055)

e. $\widehat{\Delta DHOMES_t} = -0.254 - 1.285DHOMES_{t-1}$
(se) (0.056)

f. $\widehat{\Delta DIRATE_t} = -0.0151 - 0.816DIRATE_{t-1} + 0.151\Delta DIRATE_{t-1}$
(se) (0.069) (0.058)

g. In the following test equation the $\hat{e}_t$ are the residuals from estimating the equation $HOMES_t = \beta_1 + \beta_2 IRATE_t + e_t$.

$$\widehat{\Delta\hat{e}_t} = -0.0191\hat{e}_{t-1} - 0.181\Delta\hat{e}_{t-1}$$
(se) (0.0117) (0.057)

h. In the following test equation the $\hat{u}_t$ are the residuals from estimating the equation $HOMES_t = \beta_1 + \delta_t + \beta_2 IRATE_t + u_t$.

$$\widehat{\Delta\hat{u}_t} = -0.0180\hat{u}_{t-1} - 0.208\Delta\hat{u}_{t-1}$$
(se) (0.0114) (0.057)

12.7.2 Computer Exercises

12.7 The data file *usmacro* contains quarterly observations on the U.S. unemployment rate (U), the U.S. GDP growth rate (G), and the U.S. inflation rate (INF) from 1948Q1 to 2016Q1. Plot these series and perform unit root tests on them to assess whether or not they are stationary. In your answer, justify your choice of a test equation, present the results from estimating that equation, state the null and alternative hypotheses, and draw a conclusion. Use a 5% significance level. What are the orders of integration of the three series?

12.8 The data file *okun5_aus* contains quarterly observations on the Australian unemployment rate (U), and the Australian GDP growth rate (G) from 1978Q2 to 2016Q2. Plot these series and perform unit root tests on them to assess whether or not they are stationary. In your answer, justify your choice of a test equation, present the results from estimating that equation, state the null and alternative hypotheses, and draw a conclusion. Use a 5% significance level. What are the orders of integration of the two series?

12.9 The data file *phillips5_aus* contains quarterly observations on the Australian unemployment rate (U), and the Australian inflation rate (INF) from 1987Q1 to 2016Q1. Plot these series and perform unit root tests on them to assess whether or not they are stationary. In your answer, justify your choice of a test equation, present the results from estimating that equation, state the null and alternative hypotheses, and draw a conclusion. Use a 5% significance level. What are the orders of integration of the two series?

12.10 The data file *oil5* contains quarterly observations on the price of oil from 1980Q1 to 2016Q1.

- **a.** Plot the observations.
- **b.** Using data from 1980Q1 to 2015Q2, test whether the series is stationary or nonstationary. What is its order of integration?
- **c.** Using information from part (b), the sample period 1980Q1 to 2015Q2, and other relevant criteria, specify and estimate an AR model for the price of oil.
- **d.** Use the model estimated in part (c) to forecast the price of oil for 2015Q3, 2015Q4, and 2016Q1.
- **e.** Find the percentage forecast errors for each of the forecasts made in part (d). Are your forecasts accurate?

12.11 The data file *freddie1* contains a monthly housing price index for the price of houses in Beckley, West Virginia (*BEKLY*), and the monthly value of Australian exports to China (*XCHINA*), from 1988M1 to 2015M12.

- **a.** Estimate the regression equation $XCHINA_t = \beta_1 + \beta_2\, BEKLY_t + e_t$ and comment on the results.
- **b.** Plot the series *BEKLY*, *XCHINA*, and ln(*XCHINA*) and describe the graphs. Do they provide any insights into the results from part (a)?
- **c.** Estimate the equation $\ln(XCHINA_t) = \beta_1 + \delta t + \beta_2\, BEKLY_t + e_t$ and comment on the results. Suggest a reason why ln(*XCHINA*) rather than *XCHINA* was chosen as the left-hand-side variable.
- **d.** Do unit root tests suggest ln(*XCHINA*) and *BEKLY* are stationary or trend stationary? Do the test results provide any insights into the results in part (c)?

12.12 The data file *freddie2* contains monthly housing price indices for the prices of houses in Champaign-Urbana, Illinois (*CHURB*), and Charlottesville, Virginia (*CHARV*) from 1982M1 to 2015M12.

- **a.** Plot the two series on the one graph and comment on the plots.
- **b.** Using a 5% significance level, test each of the two series for unit roots and find the order of integration of each series. Explain your choice of test equations. Are the series trend stationary? Are the series first-difference stationary? Are the series second-difference stationary?
- **c.** Using a 5% significance level, test whether *CHURB* and *CHARV* are cointegrated.
- **d.** Plot the first differences of the two series on the one graph and comment on the plots.
- **e.** Using a 5% significance level, test whether the first differences of *CHURB* and *CHARV* are cointegrated.

12.13 The data file *ozconfn* contains quarterly data on Australian real consumption expenditure (*CONS*) and real net national disposable income (*INC*) from 1975Q1 to 2010Q4.

- **a.** Create the series $LCONS = \ln(CONS)$ and plot the series *LCONS* and *INC*. Comment on the graphs.
- **b.** Detrend each of the series by estimating the linear trends $LCONS_t = a_1 + a_2 t + u_{1t}$ and $INC_t = c_1 + c_2 t + u_{2t}$, and saving the residuals. Use values $t = 0, 1, \ldots, T - 1$ for the trend term.

c. Plot the detrended series and comment on the graphs.

d. From part (c), you will have noticed that there is a strong seasonal component in each series. Econometricians have developed several methods for removing a seasonal component or "seasonally adjusting" the data. One very simple method is to subtract out the effect of seasonal dummy variables. To use this method, and remove the trend at the same time, we estimate the equation

$$y_t = \pi_0 t + \pi_1 D_{1t} + \pi_2 D_{2t} + \pi_3 D_{3t} + \pi_4 D_{4t} + u_t \tag{XR12.13}$$

where $D_{jt} = 1$ when t is an observation in quarter j, and 0 otherwise. Estimate (XR12.13) for both *LCONS* and *INC* and save the residuals; call them $LCONS^*$ and INC^*.

e. Plot $LCONS^*$ and INC^* and compare these graphs with those obtained in part (c).

f. Using a 5% significance level and the critical values in the third row of Table 12.2, test whether $LCONS^*$ and INC^* are stationary or first-difference stationary. Explain your choice of test equation, and comment on the suitability of the critical values.

g. Estimate the following two equations and compare the estimates

$$LCONS_t = \delta t + \phi_1 D_{1t} + \phi_2 D_{2t} + \phi_3 D_{3t} + \phi_4 D_{4t} + \beta INC_t + e_t$$

$$LCONS_t^* = \beta INC_t^* + e_t$$

h. Using a 5% significance level, test whether the equation in part (g)—either equation—is a cointegrating relationship. What critical value did you use?

i. Estimate an error correction model relating $\Delta LCONS_t$ to ΔINC_t and, if relevant, the lagged cointegrating residuals from part (g).

12.14 The data file *gdp5* contains the data on *GDP* displayed in Figure 12.1.

a. Is *GDP* stationary or nonstationary? Explain your choice of test equation.

b. What is the order of integration of *GDP*?

c. Construct and estimate a suitable model for forecasting *GDP* in 2017Q1. What is your forecast?

12.15 The data file *usdata5* contains the data on inflation displayed in Figure 12.1.

a. Is inflation stationary or nonstationary? Explain your choice of test equation.

b. What is the order of integration of inflation?

c. Construct and estimate a suitable model for forecasting inflation in 2017M1. What is your forecast?

12.16 In Example 12.2, using data from the data file *toody5*, we estimated the model

$$\ln(YIELD_t) = \alpha_1 + \alpha_2 t + \beta_1 RAIN_t + \beta_2 RAIN_t^2 + e_t$$

An assumption underlying this example was that $\ln(YIELD)$, *RAIN*, and $RAIN^2$ are all trend stationary. Test this assumption using a 5% significance level.

12.17 a. Using data from the data file *toody5*, estimate the following model. Comment on the results.

$$YIELD_t = \alpha_1 + \alpha_2 t + \beta_1 RAIN_t + \beta_2 RAIN_t^2 + e_t$$

b. Plot the residuals from the model estimated in part (a) and check the residual correlogram. What do you observe?

c. Estimate the following model and comment on the results.

$$YIELD_t = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \beta_1 RAIN_t + \beta_2 RAIN_t^2 + e_t$$

d. Plot the residuals from the model estimated in part (c) and check the residual correlogram. How do the properties of the residuals differ from those in part (b)?

e. Using a 5% significance level, test whether *YIELD*, *RAIN*, and $RAIN^2$ are trend stationary after subtracting out the quadratic trend.

12.18 Consider the ARDL model

$$y_t = \delta + \sum_{s=1}^{3} \theta_s y_{t-s} + \sum_{r=0}^{5} \delta_r x_{t-r} + v_t \tag{XR12.18}$$

Assume that y_t and x_t are I(1) and cointegrated. Let the cointegrating relationship be described by the equation $y_t = \beta_1 + \beta_2 x_t + e_t$.

a. Show that $\beta_1 = \delta/(1-\theta_1-\theta_2-\theta_3)$ and $\beta_2 = \sum_{r=0}^{5}\delta_r/(1-\theta_1-\theta_2-\theta_3)$.
b. Consider the corresponding error correction model

$$\Delta y_t = -\alpha(y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \sum_{r=0}^{4} \eta_r \Delta x_{t-r} + v_t$$

Show that $\delta = \alpha\beta_1$, $\theta_1 = 1-\alpha+\phi_1$, $\theta_2 = \phi_2 - \phi_1$, $\theta_3 = -\phi_2$, $\delta_0 = \eta_0$, $\delta_1 = \alpha\beta_2 - \eta_0 + \eta_1$, $\delta_2 = \eta_2 - \eta_1$, $\delta_3 = \eta_3 - \eta_2$, $\delta_4 = \eta_4 - \eta_3$, and $\delta_5 = -\eta_4$.
c. Using the data in *usdata5*, set $y_t = BR_t$ and $x_t = FFR_t$ and find least-squares estimates of the parameters in equation (XR12.18).
d. Use nonlinear least squares to estimate equation (12.32) in Example 12.9.
e. Substitute the parameter estimates of equation (12.32) obtained in part (d) into the expressions given in part (b) and compare the estimates you get with those obtained in part (c). What conclusion can you draw from this comparison?

12.19 When we estimated an error correction model for the bond and federal funds rates in Example 12.9, we estimated the coefficients of the cointegrating relationship $BR_t = \beta_1 + \beta_2 FFR_t + e_t$ at the same time as we estimated the other coefficients. Return to that example and estimate the error correction model with the cointegrating relationship replaced by the lagged residuals $\hat{e}_{t-1} = BR_{t-1} - 1.328 - 0.832FFR_{t-1}$. Compare your estimates with those obtained in Example 12.9, reported in equation (12.32).

12.20 The data file *canada6* contains monthly Canadian/U.S. exchange rates for the period 1971M1 to 2017M3. Split the observations into two sample periods—a 1971M1–1987M12 sample period and a 1988M1–2017M3 sample period.

a. Perform a unit root test on the data for each sample period. Which Dickey–Fuller tests did you use?
b. Are the results for the two sample periods consistent?
c. Perform a unit root test for the full sample 1971M1–2017M3. What is the order of integration of the data?

12.21 The data file *csi* contains the Consumer Sentiment Index (CSI) produced by the University of Michigan for the sample period 1978M1–2006M12.

a. Perform all three Dickey–Fuller tests. Are the results consistent? If not, why not?
b. Based on a graphical inspection of the data, which test should you have used?
c. Does the CSI suggest that consumers "remember" and "retain" news information for a short time, or for a long time?

12.22 The data file *mexico* contains real GDP for Mexico and the United States from the first quarter of 1980 to the third quarter of 2006. Both series have been standardized so that the average value in 2,000 is 100.

a. Perform the test for cointegration between Mexico and the Unites States using all three test equations in (12.29). Are the results consistent?
b. The theory of convergence in economic growth suggests that the two GDPs should be proportional and cointegrated. That is, there should be a cointegrating relationship that does not contain an intercept or a trend. Do your results support this theory?
c. If the variables are not cointegrated, what should you do if you are interested in testing the relationship between Mexico and the United States?

12.23 The data file *inter2* contains 300 observations of a generated I(2) process shown in Figure 12.10. Show that the variable called *inter2* is indeed an I(2) variable by conducting a number of unit root tests—first on the level of the data, then on the first difference, and finally on the second difference.

12.24 Prices around the world tend to move together. The data file *ukpi* contains information about the price indices in the United Kingdom and in the Euro Area (the United Kingdom is a member of the European Union, but not a member of the single European currency zone) for the period 1996M1–2009M12.

a. Plot the data. Are the series I(1) or I(0)?
b. Are prices in the UK and in the Euro Area cointegrated, or spuriously related? Use both the least squares and the error correction method to test this proposition.

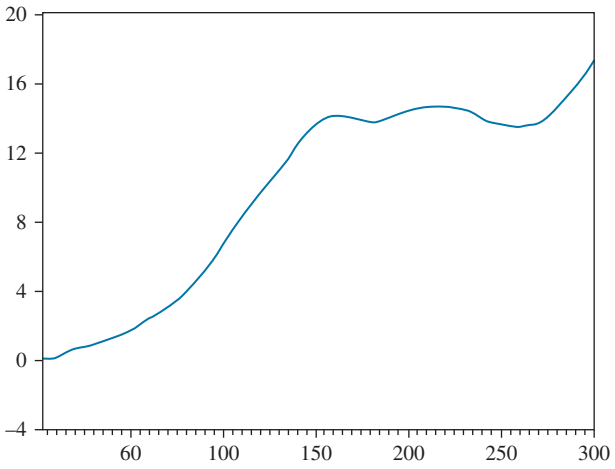


FIGURE 12.10 **A generated I(2) process.**

12.25 The data file *nasa* contains annual data on sunspots and the rate of growth in real GDP in the U.S. for the period 1950–2014. Jevons, a 19th century economist, suggested that there might be a relationship between business cycles and sunspots because variations in sunspots indicate variations in weather which in turn causes variation in agricultural output.

- **a.** Plot each of the series. Do business cycles tend to follow sunspot activity?
- **b.** Using a 5% significance level, test whether each series is stationary.
- **c.** Set up an ARDL model to test the hypothesis that sunspots can be used to predict business cycles in the U.S. Do your results support Jevons' theory?

12.26 The data file *shiller* contains the stock market data in the book "Irrational Exuberance" by Robert Shiller.[10] They comprise the monthly price and dividends of the S&P Index (in logs) for the sample period 1871M1–2015M9. Finance theory suggests a long-run relationship between dividends and the stock price.

- **a.** Plot each of the series. Do they appear to be moving together?
- **b.** Carry out an empirical analysis to investigate whether there is evidence of a long-run relationship between the two series. Use a 1% level of significance for all hypothesis tests.

12.27 How easy is it to forecast the Australian/U.S. dollar exchange rate? The data file *iron* contains monthly data on the iron ore price and the exchange rate from 2010M1 to 2016M12. In the questions that follow, use a 5% significance level for all hypothesis tests.

- **a.** Plot the two series. Do they appear to move together?
- **b.** Is the exchange rate stationary or nonstationary? What model best reflects the relationship between current and past exchange rates?
- **c.** Is the iron ore price stationary or nonstationary?
- **d.** Financial commentators have suggested that, given Australia's dependence on iron ore exports, its exchange rate follows movements in the iron ore price. Is there evidence to suggest these financial commentators are correct?
- **e.** Can the iron ore price be used to help forecast the exchange rate?

12.28 The data file *inflation* contains quarterly observations on the inflation rates in Germany and France from 1990Q1 to 2014Q4. For any hypothesis tests in the following questions, use a 5% significance level.

- **a.** Plot each of the series and comment on the plots.
- **b.** Use unit root tests, checks for serial correlation in the errors and significance of coefficients to specify and estimate an equation relating Germany's current inflation rate to its past rates.

[10]Robert Shiller, *Irrational Exuberence*, 3rd ed, 2016, Princeton University Press.

c. Use unit root tests, checks for serial correlation in the errors and significance of coefficients to specify and estimate an equation relating France's current inflation rate to its past rates.
d. Are the inflation rates in France and Germany cointegrated?
e. Specify and estimate an equation relating Germany's current exchange rate to past exchange rates in France and Germany.

12.29 Reconsider Example 6.20 where a logistic growth curve for the share of U.S. steel produced by electric arc furnace (EAF) technology was estimated. The data are stored in the data file *steel*. The curve is given by the equation

$$y_t = \frac{\alpha}{1 + \exp(-\beta - \delta t)} + e_t$$

a. Plot the series $y_t = EAF_t$. Does it give the appearance of being stationary or nonstationary? Does the logistic growth curve appear to be a good model for modeling its trend?
b. Using a 5% significance level, test the series $y_t = EAF_t$ for a unit root.
c. Estimate the equation by nonlinear least squares and plot the residuals. Do the residuals appear to be stationary. Test the residuals for a unit root.
d. Using a 5% significance level, test the series $\Delta y_t = \Delta EAF_t$ for a unit root.
e. Estimate a first-differenced version of the model and plot the residuals. Do the residuals appear to be stationary. Test the residuals for a unit root.
f. Based on your answers to the previous parts of this question, do you think $y_t = EAF_t$ is trend stationary? Compare the estimates from parts (c) and (e). Do you think the nonlinear least-squares estimates in part (c) are reliable?

CHAPTER 13

Vector Error Correction and Vector Autoregressive Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to do the following:

1. Explain why economic variables are dynamically interdependent.
2. Explain the VEC model.
3. Explain the importance of error correction.
4. Explain the VAR model.
5. Explain the relationship between a VEC model and a VAR model.
6. Explain how to estimate the VEC and VAR models for the bivariate case.
7. Explain how to generate impulse response functions and variance decompositions for the simple case when the variables are not contemporaneously interdependent and the shocks are not correlated.

KEYWORDS

dynamic relationships
error correction
forecast error variance decomposition
identification problem
impulse response functions
VAR model
VEC model

In Chapter 12, we studied the time-series properties of data and cointegrating relationships between pairs of nonstationary series. In those examples, we assumed that one of the variables was the dependent variable $\left(\text{let us call it } y_t\right)$ and that the other was the independent variable $\left(\text{say } x_t\right)$, and we treated the relationship between y_t and x_t like a regression model. However, a priori, unless we have good reasons not to, we could just as easily have assumed that y_t is the independent variable and x_t is the dependent variable. Put simply, we are working with two variables $\{y_t, x_t\}$ and the two possible regression models relating them are

$$y_t = \beta_{10} + \beta_{11} x_t + e_t^y, \quad e_t^y \sim N\left(0, \sigma_y^2\right) \tag{13.1a}$$

$$x_t = \beta_{20} + \beta_{21} y_t + e_t^x, \quad e_t^x \sim N\left(0, \sigma_x^2\right) \tag{13.1b}$$

In this bivariate (two series) system, there can be only one unique relationship between x_t and y_t, and so it must be the case that $\beta_{21} = 1/\beta_{11}$ and $\beta_{20} = -\beta_{10}/\beta_{11}$. A bit of terminology: for (13.1a), we say that we have normalized on y (meaning that the coefficient in front of y is set to 1), whereas for (13.1b), we say that we have normalized on x (meaning that the coefficient in front of x is set to 1).

Is it better to write the relationship as (13.1a) or (13.1b), or is it better to recognize that in many relationships, variables like y and x are simultaneously determined? The aim of this chapter is to explore the causal relationship between pairs of time-series variables. In doing so, we shall be extending our study of time-series data to take account of their dynamic properties and interactions. In particular, we will discuss the **vector error correction (VEC)** and **vector autoregressive (VAR)** models. We will learn how to estimate a **VEC model** when there is cointegration between I(1) variables, and how to estimate a **VAR model** when there is no cointegration. Note that this is an extension of the single-equation models examined in Chapter 12.

Some important terminology emerges here. Univariate analysis examines a single data series. Bivariate analysis examines a pair of series. The term **vector** indicates that we are considering a number of series: two, three, or more. The term "vector" is a generalization of the univariate and bivariate cases.

13.1 VEC and VAR Models

Let us begin with two time-series variables y_t and x_t and generalize the discussion about **dynamic relationships** in Chapter 9 to yield a system of equations:

$$\begin{aligned} y_t &= \beta_{10} + \beta_{11} y_{t-1} + \beta_{12} x_{t-1} + v_t^y \\ x_t &= \beta_{20} + \beta_{21} y_{t-1} + \beta_{22} x_{t-1} + v_t^x \end{aligned} \tag{13.2}$$

The equation (13.2) describes a system in which each variable is a function of its own lag and the lag of the other variable in the system. In this case, the system contains two variables y and x. In the first equation y_t is a function of its own lag y_{t-1} and the lag of the other variable in the system x_{t-1}. In the second equation x_t is a function of its own lag x_{t-1} and the lag of the other variable in the system y_{t-1}. Together the equations constitute a system known as a VAR. In this example, since the maximum lag is of order 1, we have a VAR(1).

If y and x are stationary I(0) variables, the above system can be estimated using least squares applied to each equation. If, however, y and x are nonstationary I(1) and not cointegrated, then as discussed in Chapter 12, we work with the first differences. In this case, the VAR model is

$$\begin{aligned} \Delta y_t &= \beta_{11} \Delta y_{t-1} + \beta_{12} \Delta x_{t-1} + v_t^{\Delta y} \\ \Delta x_t &= \beta_{21} \Delta y_{t-1} + \beta_{22} \Delta x_{t-1} + v_t^{\Delta x} \end{aligned} \tag{13.3}$$

All variables are now I(0), and the system can again be estimated by least squares. To recap, the VAR model is a general framework to describe the dynamic interrelationship between stationary variables. Thus, if y and x are stationary I(0) variables, the system in (13.2) is used. On the other hand, if y and x are I(1) variables but are not cointegrated, we examine the interrelation between them using a VAR framework in first differences (13.3).

If y and x are I(1) and cointegrated, then we need to modify the system of equations to allow for the cointegrating relationship between the I(1) variables. We do this for two reasons. First, as economists, we like to retain and use valuable information about the cointegrating relationship, and second, as econometricians, we like to ensure that we use the best technique that takes into account the properties of the time-series data. Recall the chapter on simultaneous equations—the cointegrating equation is one way of introducing simultaneous interactions without requiring the data to be stationary. Introducing the cointegrating relationship leads to a model known as the VEC model. We turn now to this model.

Consider two nonstationary variables y_t and x_t that are integrated of order 1: $y_t \sim \text{I}(1)$ and $x_t \sim \text{I}(1)$ and which we have shown to be cointegrated, so that

$$y_t = \beta_0 + \beta_1 x_t + e_t \tag{13.4}$$

and $\hat{e}_t \sim \text{I}(0)$ where $\hat{e}_t$ are the estimated residuals. Note that we could have chosen to normalize on x. Whether we normalize on y or x is often determined from economic theory; the critical point is that there can be at most one fundamental relationship between the two variables.

The VEC model is a special form of the VAR for I(1) variables that are cointegrated. The VEC model is

$$\begin{aligned} \Delta y_t &= \alpha_{10} + \alpha_{11}\left(y_{t-1} - \beta_0 - \beta_1 x_{t-1}\right) + v_t^y \\ \Delta x_t &= \alpha_{20} + \alpha_{21}\left(y_{t-1} - \beta_0 - \beta_1 x_{t-1}\right) + v_t^x \end{aligned} \tag{13.5a}$$

which we can expand as

$$\begin{aligned} y_t &= \alpha_{10} + \left(\alpha_{11} + 1\right)y_{t-1} - \alpha_{11}\beta_0 - \alpha_{11}\beta_1 x_{t-1} + v_t^y \\ x_t &= \alpha_{20} + \alpha_{21} y_{t-1} - \alpha_{21}\beta_0 - \left(\alpha_{21}\beta_1 - 1\right)x_{t-1} + v_t^x \end{aligned} \tag{13.5b}$$

Comparing (13.5b) with (13.2) shows the VEC as a VAR where the I(1) variable y_t is related to other lagged variables $\left(y_{t-1} \text{ and } x_{t-1}\right)$ and where the I(1) variable x_t is also related to the other lagged variables $\left(y_{t-1} \text{ and } x_{t-1}\right)$. Note, however, that the two equations contain the common cointegrating relationship.

The coefficients α_{11}, α_{21} are known as **error correction** coefficients, so named because they show how much Δy_t and Δx_t respond to the cointegrating error $y_{t-1} - \beta_0 - \beta_1 x_{t-1} = e_{t-1}$. The idea that the error leads to a correction comes about because of the conditions put on α_{11}, α_{21} to ensure stability, namely $\left(-1 < \alpha_{11} \leq 0\right)$ and $\left(0 \leq \alpha_{21} < 1\right)$. To appreciate this idea, consider a positive error $e_{t-1} > 0$ that occurred because $y_{t-1} > \left(\beta_0 + \beta_1 x_{t-1}\right)$. A negative error correction coefficient in the first equation $\left(\alpha_{11}\right)$ ensures that Δy falls, while the positive error correction coefficient in the second equation $\left(\alpha_{21}\right)$ ensures that Δx rises, thereby correcting the error. Having the error correction coefficients less than 1 in absolute value ensures that the system is not explosive. Note that the VEC is a generalization of the error-correction (single-equation) model discussed in Chapter 12. In the VEC (system) model, both y_t and x_t "error-correct."

The error correction model has become an extremely popular model because its interpretation is intuitively appealing. Think about two nonstationary variables, say consumption (let us call it y_t) and income (let us call it x_t), that we expect to be related (cointegrated). Now think about a change in your income Δx_t, say a pay raise! Consumption will most likely increase, but it may take you a while to change your consumption pattern in response to a change in your pay. The VEC model allows us to examine how much consumption will change in response to a change in the explanatory variable $\left(\text{the cointegration part, } y_t = \beta_0 + \beta_1 x_t + e_t\right)$, as well as the speed of the change (the error correction part, $\Delta y_t = \alpha_{10} + \alpha_{11}\left(e_{t-1}\right) + v_t^y$ where e_{t-1} is the cointegrating error).

There is one final point to discuss—the role of the intercept terms. Thus far, we have introduced an intercept term in the cointegrating equation (β_0) as well as in the VEC $\left(\alpha_{10} \text{ and } \alpha_{20}\right)$. However, doing so can create a problem. To see why, we collect all the intercept terms and rewrite (13.5b) as

$$\begin{aligned} y_t &= \left(\alpha_{10} - \alpha_{11}\beta_0\right) + \left(\alpha_{11} + 1\right)y_{t-1} - \alpha_{11}\beta_1 x_{t-1} + v_t^y \\ x_t &= \left(\alpha_{20} - \alpha_{21}\beta_0\right) + \alpha_{21} y_{t-1} - \left(\alpha_{21}\beta_1 - 1\right)x_{t-1} + v_t^x \end{aligned} \tag{13.5c}$$

If we estimate each equation by least squares, we obtain estimates of composite terms $\left(\alpha_{10} - \alpha_{11}\beta_0\right)$ and $\left(\alpha_{20} - \alpha_{21}\beta_0\right)$, and we are not able to disentangle the separate effects of β_0, α_{10}, and α_{20}. In the next section, we discuss a simple two-step least squares procedure that gets around this problem. However, the lesson here is to check whether, and where, an intercept term is needed.

13.2 Estimating a Vector Error Correction Model

There are many econometric methods to estimate the error correction model. Nonlinear (system) least squares is one method, but the most straightforward method is to use a two-step least squares procedure. First, use OLS to estimate the cointegrating relationship $y_t = \beta_0 + \beta_1 x_t + e_t$ and generate the lagged residuals $\hat{e}_{t-1} = y_{t-1} - b_0 - b_1 x_{t-1}$.

Second, use OLS to estimate the equations:

$$\Delta y_t = \alpha_{10} + \alpha_{11}\hat{e}_{t-1} + v_t^y \tag{13.6a}$$

$$\Delta x_t = \alpha_{20} + \alpha_{21}\hat{e}_{t-1} + v_t^x \tag{13.6b}$$

Note that all the variables in (13.6) (Δy, Δx, and $\hat{e}$) are stationary (recall that for y and x to be cointegrated, the residuals $\hat{e}$ must be stationary). Hence, the standard regression analysis studied in earlier chapters may be used to test the significance of the parameters. The usual residual diagnostic tests may be applied.

We need to be careful here about how we combine stationary and nonstationary variables in a regression model. Cointegration is about the relationship between I(1) variables. The cointegrating equation does not contain I(0) variables. The corresponding VEC model, however, relates the change in an I(1) variable (the I(0) variables Δy and Δx) to other I(0) variables, namely, the cointegration residuals $\hat{e}_{t-1}$; if required, other stationary variables may be added. In other words, we should not mix stationary and nonstationary variables: an I(0) dependent variable on the left-hand side of a regression equation should be "explained" by other I(0) variables on the right-hand side and an I(1) dependent variable on the left-hand side of a regression equation should be explained by other I(1) variables on the right-hand side.

EXAMPLE 13.1 | VEC Model for GDP

In Figure 13.1 the quarterly real gross domestic product (GDP) of a small economy (Australia) and a large economy (United States) for the sample period 1970Q1 to 2000Q4 are displayed. Note that the series have been scaled so that both economies show a real GDP value of 100 in 2000. They appear in the data file *gdp*. It appears from the figure that both series are nonstationary and possibly cointegrated.

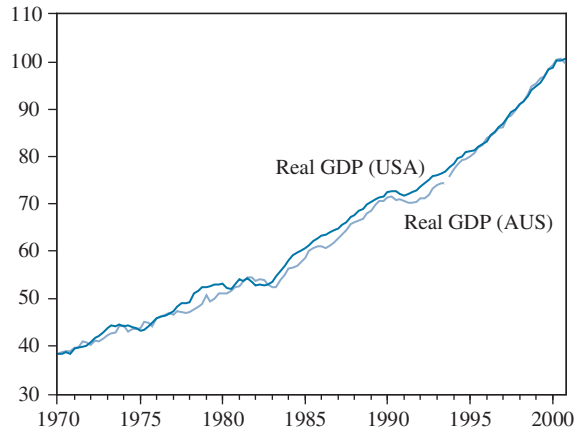


FIGURE 13.1 **Real gross domestic product (GDP = 100 in 2000).**

Formal unit root tests of the series confirm that they are indeed nonstationary. To check for cointegration we obtain the fitted equation in (13.7) (the intercept term is omitted because it has no economic meaning):

$$\hat{A}_t = 0.985 U_t, \tag{13.7}$$

where A denotes real GDP for Australia and U denotes real GDP for the United States. Note that we have normalized on A because it makes more sense to think of a small economy responding to a large economy. The residuals derived from the cointegrating relationship $\hat{e}_t = A_t - 0.985 U_t$ are shown in Figure 13.2. Their first-order autocorrelation is 0.870, and a visual inspection of the time series suggests that the residuals may be stationary.

A formal unit root test is performed, and the estimated unit root test equation is

$$\widehat{\Delta e_t} = -0.128\hat{e}_{t-1} \tag{13.8}$$
$$(tau) \quad (-2.889)$$

Since the cointegrating relationship does not contain an intercept term [see Chapter 12, (12.29a)], the 5% critical value is −2.76. The unit root t-value of −2.889 is less than −2.76. We reject the null of no cointegration and we conclude that the two real GDP series are cointegrated. This result implies that economic activity in the small economy (Australia, A_t) is linked to economic activity in the large economy (United States, U_t). If U_t were to increase by one unit, A_t would increase by 0.985. But the Australian economy may not respond fully by this amount within the quarter. To ascertain how much it will respond within a quarter, we estimate the error correction model by least squares. The estimated VEC model for $\{A_t, U_t\}$ is

$$\begin{aligned} \widehat{\Delta A_t} &= 0.492 - 0.099\hat{e}_{t-1} \\ (t) \qquad & (-2.077) \\ \widehat{\Delta U_t} &= 0.510 + 0.030\hat{e}_{t-1} \\ (t) \qquad & (0.789) \end{aligned} \tag{13.9}$$

The results show that both error correction coefficients are of the appropriate sign. The negative error correction coefficient in the first equation (−0.099) indicates that ΔA falls (i.e., A_t falls or ΔA_t is negative), while the positive error correction coefficient in the second equation (0.030) indicates that ΔU rises (i.e., U_t rises or ΔU_t is positive), when there is a positive cointegrating error $\left(\hat{e}_{t-1} > 0 \text{ or } A_{t-1} > 0.985U_{t-1}\right)$. This behavior (negative change in A and positive change in U) "corrects" the cointegrating error. The error correction coefficient (−0.099) is significant at the 5% level; it indicates that the quarterly adjustment of A_t will be about 10% of the deviation of A_{t-1} from its cointegrating value $0.985U_{t-1}$. This is a slow rate of adjustment. However, the error correction coefficient in the second equation (0.030) is insignificant; it suggests that ΔU does not react to the cointegrating error. This outcome is consistent with the view that the small economy is likely to react to economic conditions in the large economy, but not vice versa.

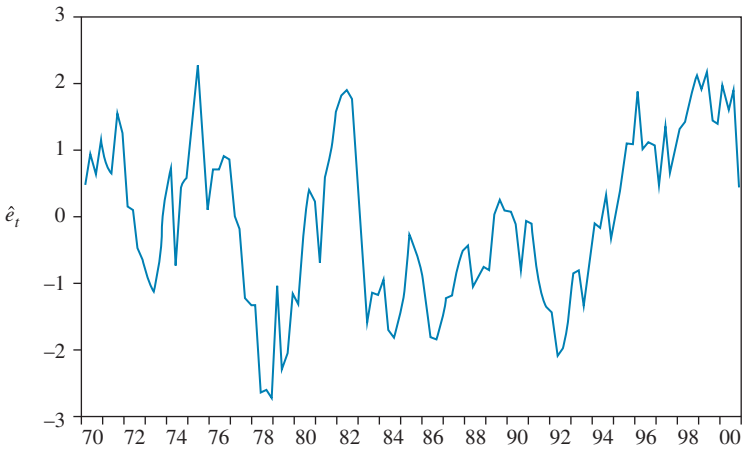


FIGURE 13.2 **Residuals derived from the cointegrating relationship.**

13.3 Estimating a VAR Model

The VEC is a multivariate dynamic model that incorporates a cointegrating equation. It is relevant when, for the bivariate case, we have two variables, say y and x, that are both I(1), but are cointegrated. Now we ask: What should we do if we are interested in the interdependencies between y and x, but they are not cointegrated? In this case, we estimate a VAR model as shown in (13.3).

EXAMPLE 13.2 | VAR Model for Consumption and Income

Consider Figure 13.3 that shows the log of real personal disposable income (RPDI) (denoted as Y) and the log of real personal consumption expenditure (RPCE) (denoted as C) for the U.S. economy over the period 1986Q1 to 2015Q2. Both series appear to be nonstationary, but are they cointegrated? The quarterly data are stored in the data file *fred5*.

The Dickey–Fuller test values for unit roots for C were -0.88 when an intercept only was included and -1.63 when both an intercept and trend term were included. In both cases, there were three augmentation terms. The corresponding values for Y were -1.65 and -0.43. In these cases, one augmentation term was sufficient. The 10% critical values from Table 12.2 are -2.57 without a trend and -3.13 with a trend. Since the test values are greater than the critical values, we cannot conclude that the series are stationary. Using a 10% significance level, unit root tests on the first differences of the series lead to a conclusion that the first differences are stationary, and hence the series are I(1). Testing for cointegration yields the following results:

$$\begin{aligned} \hat{e}_t &= C_t + 0.543 - 1.049Y_t \\ \widehat{\Delta\hat{e}_t} &= -0.203\hat{e}_{t-1} - 0.290\Delta\hat{e}_{t-1} \\ (\tau) \quad & (-3.046) \end{aligned} \tag{13.10}$$

An intercept term has been included to capture the component of (log) consumption that is independent of disposable income. From Table 12.4, the 10% critical value of the test for stationarity in the cointegrating residuals is -3.07. Since the *tau* (unit root t-value) of -3.046 is greater than -3.07, it indicates that the errors are not stationary and hence that the relationship between C (i.e., log(RPCE)) and Y (i.e., log(RPDI)) is spurious. That is, we have no cointegration. Thus, we would not apply a VEC model to examine the dynamic relationship between aggregate consumption C and income Y. Instead, we would estimate a VAR model for the set of I(0) variables $\{\Delta C_t, \Delta Y_t\}$.

For illustrative purposes, the order of lag in this example has been restricted to one. In general, one should use significance of the coefficient estimates and serial correlation in the errors to choose a suitable number of lags which may be greater than one. The results are

$$\widehat{\Delta C_t} = 0.00367 + 0.348\Delta C_{t-1} + 0.131\Delta Y_{t-1}$$
$$(t) \quad (4.87) \quad (4.02) \quad (2.52) \tag{13.11a}$$

$$\widehat{\Delta Y_t} = 0.00438 + 0.590\Delta C_{t-1} - 0.291\Delta Y_{t-1}$$
$$(t) \quad (3.38) \quad (3.96) \quad (-3.25) \tag{13.11b}$$

The first equation (13.11a) shows that the quarterly growth in consumption (ΔC_t) is significantly related to its own past value (ΔC_{t-1}) and also significantly related to the quarterly growth in last period's income (ΔY_{t-1}). The second equation (13.11b) shows that ΔY_t is significantly negatively related to its own past value but significantly positively related to last period's change in consumption. The constant terms capture the fixed component in the change in log consumption and the change in log income.

Having estimated these models, can we infer anything else? If the system is subjected to an income shock, what is the effect of the shock on the dynamic path of the quarterly growth in consumption and income? Will they rise and by how much? If the system is also subjected to a consumption shock, what is the contribution of an income versus a consumption shock on the variation of income? We turn now to some analysis suited to addressing these questions.

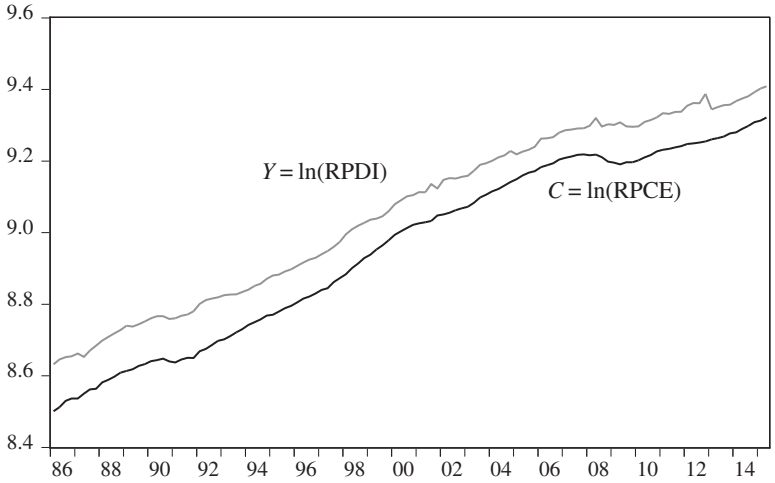


FIGURE 13.3 **The logarithms of real personal disposable income (RPDI) and real personal consumption expenditure (RPCE).**

13.4 Impulse Responses and Variance Decompositions

Impulse response functions and variance decompositions are techniques that are used by macroeconometricians to analyze problems such as the effect of an oil price shock on inflation and GDP growth, and the effect of a change in monetary policy on the economy.

13.4.1 Impulse Response Functions

Impulse response functions show the effects of shocks on the adjustment path of the variables. To help us understand this, we shall first consider a univariate series.

The Univariate Case

Consider a univariate series $y_t = \rho y_{t-1} + v_t$ and subject it to a shock of size v in period one. Assume an arbitrary starting value of y at time zero: $y_0 = 0$. (Since we are interested in the dynamic path, the starting point is irrelevant.) At time $t = 1$, following the shock, the value of y will be: $y_1 = \rho y_0 + v_1 = v$. Assume that there are no subsequent shocks in later time periods $\left[v_2 = v_3 = \cdots = 0\right]$, at time $t = 2$, $y_2 = \rho y_1 = \rho v$. At time $t = 3$, $y_3 = \rho y_2 = \rho\left(\rho y_1\right) = \rho^2 v$, and so on. Thus the time-path of y following the shock is $\left\{v, \rho v, \rho^2 v, \ldots\right\}$. The values of the coefficients $\left\{1, \rho, \rho^2, \ldots\right\}$ are known as multipliers, and the time-path of y following the shock is known as the impulse response function.

To illustrate, assume that $\rho = 0.9$ and let the shock be unity: $v = 1$. According to the analysis, y will be $\{1, 0.9, 0.81, \ldots\}$, approaching zero over time. This impulse response function is plotted in Figure 13.4. It shows us what happens to y after a shock. In this case, y initially rises by the full amount of the shock and then it gradually returns to the value before the shock.

The Bivariate Case

Now, let us consider an impulse response function analysis with two time series based on a bivariate VAR system of stationary variables:

$$\begin{aligned} y_t &= \delta_{10} + \delta_{11} y_{t-1} + \delta_{12} x_{t-1} + v_t^y \\ x_t &= \delta_{20} + \delta_{21} y_{t-1} + \delta_{22} x_{t-1} + v_t^x \end{aligned} \tag{13.12}$$

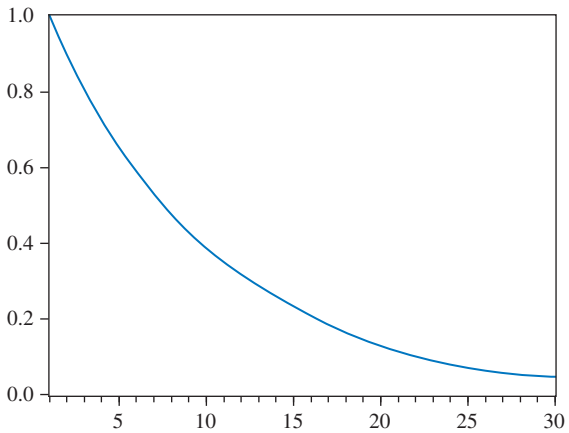


FIGURE 13.4 **Impulse responses for an AR(1) model $y_t = 0.9\, y_{t-1} + v_t$ following a unit shock.**

In this case, there are two possible shocks to the system—one to y and the other to x. Thus we are interested in four impulse response functions—the effect of a shock to y on the time-paths of y and x and the effect of a shock to x on the time-paths of y and x.

The actual mechanics of generating impulse responses in a system is complicated by the facts that (i) one has to allow for interdependent dynamics (the multivariate analog of generating the multipliers) and (ii) one has to identify the correct shock from unobservable data. Taken together, these two complications lead to what is known as the **identification problem**. In this chapter, we consider a special case where there is no identification problem.[1] This special case occurs when the system that is described in (13.12) is a true representation of the dynamic system—namely, y is related only to lags of y and x, and x is related only to lags of y and x. In other words, y and x are related dynamically, but not contemporaneously. The current value x_t does not appear in the equation for y_t and the current value y_t does not appear in the equation for x_t. Also, we need to assume that the errors v_t^x and v_t^y are contemporaneously uncorrelated.

Consider the case when there is a one standard deviation shock (alternatively called an **innovation**) to y so that at time $t = 1$, $v_1^y = \sigma_y$, and v_t^y is zero thereafter. Assume $v_t^x = 0$ for all t. It is traditional to consider a standard deviation shock (innovation) rather than a unit shock to eliminate units of measurement. Assume $y_0 = x_0 = 0$. Also, since we are focusing on how a shock *changes* the paths of y and x, we can ignore the intercepts. Then

1. When $t = 1$, the effect of a shock of size σ_y on y is $y_1 = v_1^y = \sigma_y$, and the effect on x is $x_1 = v_1^x = 0$.
2. When $t = 2$, the effect of the shock on y is

$$y_2 = \delta_{11}y_1 + \delta_{12}x_1 = \delta_{11}\sigma_y + \delta_{12}0 = \delta_{11}\sigma_y$$

and the effect on x is

$$x_2 = \delta_{21}y_1 + \delta_{22}x_1 = \delta_{21}\sigma_y + \delta_{22}0 = \delta_{21}\sigma_y$$

3. When $t = 3$, the effect of the shock on y is

$$y_3 = \delta_{11}y_2 + \delta_{12}x_2 = \delta_{11}\delta_{11}\sigma_y + \delta_{12}\delta_{21}\sigma_y$$

and the effect on x is

$$x_3 = \delta_{21}y_2 + \delta_{22}x_2 = \delta_{21}\delta_{11}\sigma_y + \delta_{22}\delta_{21}\sigma_y.$$

By repeating the substitutions for $t = 4, 5, \ldots$, we obtain further expressions. The impulse response of the shock (or innovation) to y on y is $\sigma_y\left[1, \delta_{11}, \left(\delta_{11}\delta_{11} + \delta_{12}\delta_{21}\right), \ldots\right]$ and the impulse response of a shock to y on x is $\sigma_y\left[0, \delta_{21}, \left(\delta_{21}\delta_{11} + \delta_{22}\delta_{21}\right), \ldots\right]$.

Now consider what happens when there is a one standard deviation shock to x so that at time $t = 1$, $v_1^x = \sigma_x$, and v_t^x is zero thereafter. Assume $v_t^y = 0$ for all t. In the first period after the shock, the effect of a shock of size σ_x on y is $y_1 = v_1^y = 0$, and the effect of the shock on x is $x_1 = v_1^x = \sigma_x$. Two periods after the shock, when $t = 2$, the effect on y is

$$y_2 = \delta_{11}y_1 + \delta_{12}x_1 = \delta_{11}0 + \delta_{12}\sigma_x = \delta_{12}\sigma_x$$

and the effect on x is

$$x_2 = \delta_{21}y_1 + \delta_{22}x_1 = \delta_{21}0 + \delta_{22}\sigma_x = \delta_{22}\sigma_x$$

Again, by repeated substitutions, we obtain the impulse response of a shock to x on y as $\sigma_x\left[0, \delta_{12}, \left(\delta_{11}\delta_{12} + \delta_{12}\delta_{22}\right), \ldots\right]$, and the impulse response of a shock to x on x as $\sigma_x\left[1, \delta_{22}, \left(\delta_{21}\delta_{12} + \delta_{22}\delta_{22}\right), \ldots\right]$. Figure 13.5 shows the four impulse response functions for numerical values: $\sigma_y = 1$, $\sigma_x = 2$, $\delta_{11} = 0.7$, $\delta_{12} = 0.2$, $\delta_{21} = 0.3$ and $\delta_{22} = 0.6$.

[1] Appendix 13A introduces the general problem.

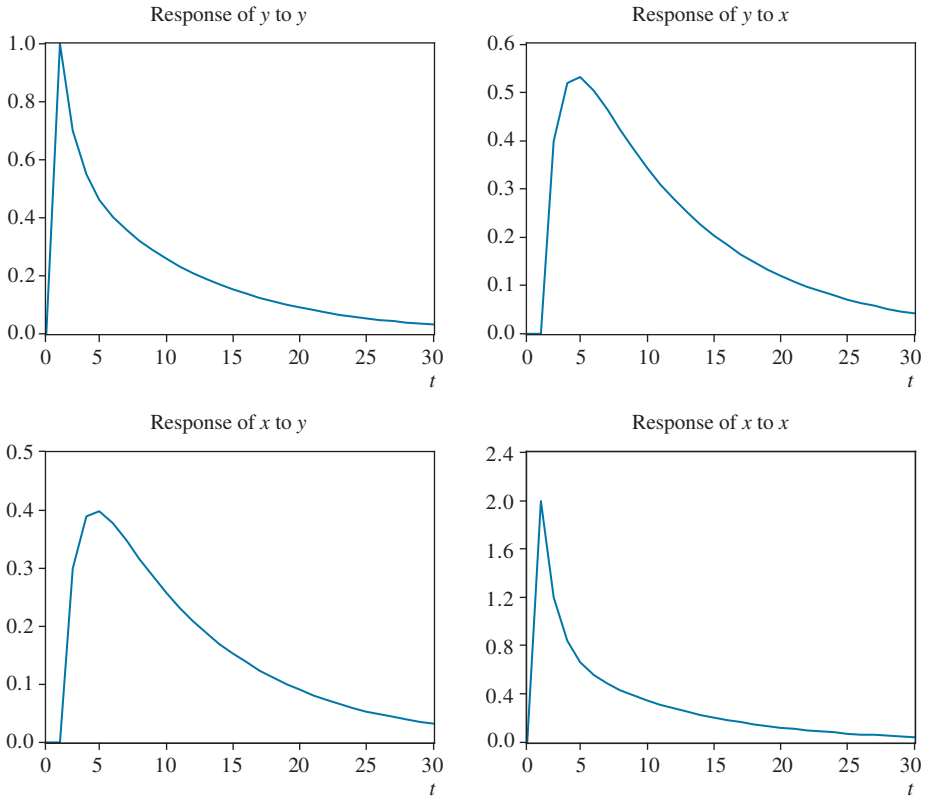


FIGURE 13.5 **Impulse responses to standard deviation shock.**

The advantage of examining impulse response functions (and not just VAR coefficients) is that they show the size of the impact of the shock plus the rate at which the shock dissipates, allowing for interdependencies.

13.4.2 Forecast Error Variance Decompositions

Another way to disentangle the effects of various shocks is to consider the contribution of each type of shock to the forecast error variance.

Univariate Analysis Consider again the univariate series, $y_t = \rho y_{t-1} + v_t$. The best one-step-ahead forecast (alternatively the forecast one period ahead) is

$$y_{t+1}^F = E_t[\rho y_t + v_{t+1}]$$

where E_t is the expected value conditional on information at time t (i.e., we are interested in the mean value of y_{t+1} using what is known at time t). At time t the conditional expectation $E_t[\rho y_t] = \rho y_t$ is known, but the error v_{t+1} is unknown, and so its conditional expectation is zero. Thus the best forecast of y_{t+1} is ρy_t, and the forecast error is

$$y_{t+1} - E_t[y_{t+1}] = y_{t+1} - \rho y_t = v_{t+1}$$

The variance of the one-step forecast error is $\text{var}(v_{t+1}) = \sigma^2$. Suppose we wish to forecast two steps ahead; using the same logic, the two-step forecast becomes

$$y_{t+2}^F = E_t[\rho y_{t+1} + v_{t+2}] = E_t[\rho(\rho y_t + v_{t+1}) + v_{t+2}] = \rho^2 y_t$$

and the two-step forecast error becomes

$$y_{t+2} - E_t[y_{t+2}] = y_{t+2} - \rho^2 y_t = \rho v_{t+1} + v_{t+2}$$

In this case, the variance of the forecast error is $\text{var}(\rho v_{t+1} + v_{t+2}) = \sigma^2(\rho^2 + 1)$, showing that the variance of forecast error increases as we increase the forecast horizon. There is only one shock that leads to a forecast error. Hence the forecast error variance is 100% due to its own shock. The exercise of attributing the source of the variation in the forecast error is known as variance decomposition.

Bivariate Analysis We can perform a **forecast error variance decomposition** for our special bivariate example where there is no identification problem. Ignoring the intercepts (since they are constants), the one-step ahead forecasts are

$$y_{t+1}^F = E_t\left[\delta_{11}y_t + \delta_{12}x_t + v_{t+1}^y\right] = \delta_{11}y_t + \delta_{12}x_t$$

$$x_{t+1}^F = E_t\left[\delta_{21}y_t + \delta_{22}x_t + v_{t+1}^x\right] = \delta_{21}y_t + \delta_{22}x_t$$

The corresponding one-step-ahead forecast errors and variances are

$$FE_1^y = y_{t+1} - E_t\left[y_{t+1}\right] = v_{t+1}^y \qquad \text{var}\left(FE_1^y\right) = \sigma_y^2$$

$$FE_1^x = x_{t+1} - E_t\left[x_{t+1}\right] = v_{t+1}^x \qquad \text{var}\left(FE_1^x\right) = \sigma_x^2$$

Hence in the first period, all variation in the forecast error for y is due to its own shock. Likewise, 100% of the forecast error for x can be explained by its own shock. Using the same technique, the two-step ahead forecast for y is

$$\begin{aligned} y_{t+2}^F &= E_t\left[\delta_{11}y_{t+1} + \delta_{12}x_{t+1} + v_{t+2}^y\right] \\ &= E_t\left[\delta_{11}(\delta_{11}y_t + \delta_{12}x_t + v_{t+1}^y) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t + v_{t+1}^x) + v_{t+2}^y\right] \\ &= \delta_{11}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{12}(\delta_{21}y_t + \delta_{22}x_t) \end{aligned}$$

and that for x is

$$\begin{aligned} x_{t+2}^F &= E_t\left[\delta_{21}y_{t+1} + \delta_{22}x_{t+1} + v_{t+2}^x\right] \\ &= E_t\left[\delta_{21}(\delta_{11}y_t + \delta_{12}x_t + v_{t+1}^y) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t + v_{t+1}^x) + v_{t+2}^x\right] \\ &= \delta_{21}(\delta_{11}y_t + \delta_{12}x_t) + \delta_{22}(\delta_{21}y_t + \delta_{22}x_t) \end{aligned}$$

The corresponding two-step-ahead forecast errors and variances are (recall that we are working with the special case of independent errors)

$$FE_2^y = y_{t+2} - E_t\left[y_{t+2}\right] = \left[\delta_{11}v_{t+1}^y + \delta_{12}v_{t+1}^x + v_{t+2}^y\right]$$

$$\text{var}\left(FE_2^y\right) = \delta_{11}^2\sigma_y^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2$$

$$FE_2^x = x_{t+2} - E_t\left[x_{t+2}\right] = \left[\delta_{21}v_{t+1}^y + \delta_{22}v_{t+1}^x + v_{t+2}^x\right]$$

$$\text{var}\left(FE_2^x\right) = \delta_{21}^2\sigma_y^2 + \delta_{22}^2\sigma_x^2 + \sigma_x^2$$

We can decompose the total variance of the forecast error for y, $\left(\delta_{11}^2\sigma_y^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2\right)$, into that due to shocks to y, $\left(\delta_{11}^2\sigma_y^2 + \sigma_y^2\right)$, and that due to shocks to x, $\left(\delta_{12}^2\sigma_x^2\right)$. This decomposition is often expressed in proportional terms. The proportion of the two-step forecast error variance of y explained by its "own" shock is

$$\left(\delta_{11}^2\sigma_y^2 + \sigma_y^2\right) \Big/ \left(\delta_{11}^2\sigma_y^2 + \delta_{12}^2\sigma_x^2 + \sigma_y^2\right)$$

and the proportion of the two-step forecast error variance of y explained by the "other" shock is

$$\left(\delta_{12}^2\sigma_x^2\right)\Big/\left(\delta_{11}^2\sigma_y^2+\delta_{12}^2\sigma_x^2+\sigma_y^2\right)$$

Similarly, the proportion of the two-step forecast error variance of x explained by its own shock is

$$\left(\delta_{22}^2\sigma_x^2+\sigma_x^2\right)\Big/\left(\delta_{21}^2\sigma_y^2+\delta_{22}^2\sigma_x^2+\sigma_x^2\right)$$

and the proportion of the forecast error of x explained by the other shock is

$$\left(\delta_{21}^2\sigma_y^2\right)\Big/\left(\delta_{21}^2\sigma_y^2+\delta_{22}^2\sigma_x^2+\sigma_x^2\right)$$

For our numerical example with $\sigma_y = 1$, $\sigma_x = 2$, $\delta_{11} = 0.7$, $\delta_{12} = 0.2$, $\delta_{21} = 0.3$, and $\delta_{22} = 0.6$, we find that 90.303% of the two-step forecast error variance of y is due to y, and only 9.697% is due to x.

To sum up, suppose you were interested in the relationship between economic growth and inflation. A VAR model will tell you whether they are significantly related to each other; an impulse response analysis will show how growth and inflation react dynamically to shocks, and a variance decomposition analysis will be informative about the sources of volatility.

The General Case The example above assumes that x and y are not contemporaneously related and that the shocks are uncorrelated. There is no identification problem, and the generation and interpretation of the impulse response functions and decomposition of the forecast error variance are straightforward. In general, this is unlikely to be the case. Contemporaneous interactions and correlated errors complicate the identification of the nature of shocks and hence the interpretation of the impulses and decomposition of the causes of the forecast error variance. This topic is discussed in greater detail in textbooks devoted to time-series analysis.[2] A description of how the identification problem can arise is given in Appendix 13A.

13.5 Exercises

13.5.1 Problems

13.1 Consider the following first-order VAR model of stationary variables:

$$y_t = \delta_{11}y_{t-1} + \delta_{12}x_{t-1} + v_t^y$$
$$x_t = \delta_{21}y_{t-1} + \delta_{22}x_{t-1} + v_t^x$$

Under the assumption that there is no contemporaneous dependence, determine the impulse responses, four periods after a standard deviation shock for

a. y following a shock to y
b. y following a shock to x
c. x following a shock to y
d. x following a shock to x

13.2 Consider the first-order VAR model in Exercise 13.1. Under the assumption that there is no contemporaneous dependence, determine

a. the contribution of a shock to y on the variance of the three-step ahead forecast error for y
b. the contribution of a shock to x on the variance of the three-step ahead forecast error for y
c. the contribution of a shock to y on the variance of the three-step ahead forecast error for x
d. the contribution of a shock to x on the variance of the three-step ahead forecast error for x

[2]One reference you might consider is Lütkepohl, H. (2005) *Introduction to Multiple Time Series Analysis*, Springer, New York, Chapter 9.

13.3 The VEC model is a special form of the VAR for I(1) variables that are cointegrated. Consider the following VEC model:

$$\Delta y_t = \alpha_{10} + \alpha_{11}(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + v_t^y$$
$$\Delta x_t = \alpha_{20} + \alpha_{21}(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + v_t^x$$

The VEC model may also be rewritten as a VAR, but the two equations will contain common parameters:

$$y_t = \alpha_{10} + (\alpha_{11} + 1)y_{t-1} - \alpha_{11}\beta_0 - \alpha_{11}\beta_1 x_{t-1} + v_t^y$$
$$x_t = \alpha_{20} + \alpha_{21}y_{t-1} - \alpha_{21}\beta_0 - (\alpha_{21}\beta_1 - 1)x_{t-1} + v_t^x$$

a. Suppose you were given the following results from an estimated VEC model:

$$\widehat{\Delta y_t} = 2 - 0.5(y_{t-1} - 1 - 0.7x_{t-1})$$
$$\widehat{\Delta x_t} = 3 + 0.3(y_{t-1} - 1 - 0.7x_{t-1})$$

Rewrite the model in the VAR form.

b. Now suppose you were given the following results of an estimated VAR model, but you were also told that y and x are cointegrated.

$$\hat{y}_t = 0.7y_{t-1} + 0.3 + 0.24x_{t-1}$$
$$\hat{x}_t = 0.6y_{t-1} - 0.6 + 0.52x_{t-1}$$

Rewrite the model in the VEC form.

13.4 VAR and VEC models are popular forecasting models because they rely on the past history of observed outcomes to predict the expected future values.

a. Consider the following estimated VAR model:

$$y_t = \hat{\delta}_{11}y_{t-1} + \hat{\delta}_{12}x_{t-1} + \hat{v}_{1t}$$
$$x_t = \hat{\delta}_{21}y_{t-1} + \hat{\delta}_{22}x_{t-1} + \hat{v}_{2t}$$

What are the forecasts for y_{t+1} and x_{t+1}?
What are the forecasts for y_{t+2} and x_{t+2}?

b. Consider the following estimated VEC model:

$$\Delta y_t = \hat{\alpha}_{11}\left(y_{t-1} - \hat{\beta}_1 x_{t-1}\right) + \hat{v}_{1t}$$
$$\Delta x_t = \hat{\alpha}_{21}\left(y_{t-1} - \hat{\beta}_1 x_{t-1}\right) + \hat{v}_{2t}$$

What are the forecasts for y_{t+1} and x_{t+1}?
What are the forecasts for y_{t+2} and x_{t+2}?

13.5.2 Computer Exercises

13.5 The data file *gdp* contains quarterly data on the real GDP of Australia (*AUS*) and real GDP of the United States (*USA*) for the sample period 1970Q1 to 2000Q4.

a. Are the series stationary or nonstationary?

b. Test for cointegration allowing for an intercept term. You will find that the intercept is negative. Is this sensible? If not, repeat the test for cointegration excluding the constant term.

c. Save the cointegrating residuals and estimate the VEC model.

13.6 The data file *fred5* contains the log of RPDI (Y) and the log of RPCE (C) for the U.S. economy over the period 1986Q1 to 2015Q2.

a. Are the series stationary, or nonstationary? In particular, test whether the series are trend stationary.

b. Test for cointegration allowing for an intercept term. Are the series cointegrated?

c. Estimate a VAR model for the set of I(0) variables $\{\Delta C_t, \Delta Y_t\}$. Pay particular attention to the order of lags.

13.7 Consider again the data file *fred5* used in Example 13.2 and Exercise 13.6.

a. Estimate a VAR model for $\{\Delta C_t, \Delta Y_t\}$ with three lags of each variable included. Comment on the results. Has serial correlation in the errors been eliminated?

b. The concept of "Granger causality" was introduced in Section 9.3.4. In a VAR involving two variables x and y, we can ask whether x Granger causes y, whether y Granger causes x, and whether there is Granger causality in both directions. Using the model estimated in part (a), test whether ΔY Granger causes ΔC and whether ΔC Granger causes ΔY.

13.8 The data file *vec* contains 100 observations on two generated series of data, x and y. The variables are nonstationary and cointegrated without a constant term. Save the cointegrating residuals $(\hat{e})$ and estimate the VEC model. As a check, the results for the case normalized on y are

$$\begin{aligned}\widehat{\Delta y_t} &= -0.576(\hat{e}_{t-1})\\ (t) \quad &(-6.158)\\ \widehat{\Delta x_t} &= 0.450(\hat{e}_{t-1})\\ (t) \quad &(4.448)\end{aligned}$$

a. The residuals from the error correction model should not be autocorrelated. Are they?

b. Note that one of the error correction terms is negative and the other is positive. Explain why this is necessary.

13.9 The data file *var* contains 100 observations on two generated series of data, w and z. The variables are nonstationary but not cointegrated. Estimate a VAR model of changes in the variables. As a check, the results are (the intercept terms were not significant):

$$\begin{aligned}\widehat{\Delta w_t} &= 0.743\Delta w_{t-1} + 0.214\Delta z_{t-1}\\ (t) \quad &(11.403) \qquad (2.893)\\ \widehat{\Delta z_t} &= -0.155\Delta w_{t-1} + 0.641\Delta z_{t-1}\\ (t) \quad &(-2.293) \qquad (8.338)\end{aligned}$$

a. The residuals from the VAR model should not be autocorrelated. Is this the case?

b. Determine the impulse responses for the first two periods. (You may assume the special condition that there is no contemporaneous dependence.)

c. Determine the variance decompositions for the first two periods.

13.10 The quantity theory of money says that there is a direct relationship between the quantity of money in the economy and the aggregate price level. Put simply, if the quantity of money doubles, then the price level should also double. Figure 13.6 shows the percentage change in a measure of the quantity of money (M) and the percentage change in a measure of aggregate prices (P) for the United States between 1961Q1 and 2005Q4 (data file *qtm*). A VEC model was estimated as follows:

$$\begin{aligned}\widehat{\Delta P_t} &= -0.016(P_{t-1} - 1.004M_{t-1} + 0.039) + 0.514\Delta P_{t-1} - 0.005\Delta M_{t-1}\\ (t) \quad &(-2.127) \quad (-3.696) \quad (1.714) \quad (7.999) \quad (-0.215)\\ \widehat{\Delta M_t} &= 0.067(P_{t-1} - 1.004M_{t-1} + 0.039) - 0.336\Delta P_{t-1} - 0.340\Delta M_{t-1}\\ (t) \quad &(3.017) \quad (-3.696) \quad (1.714) \quad (-1.796) \quad (-4.802)\end{aligned}$$

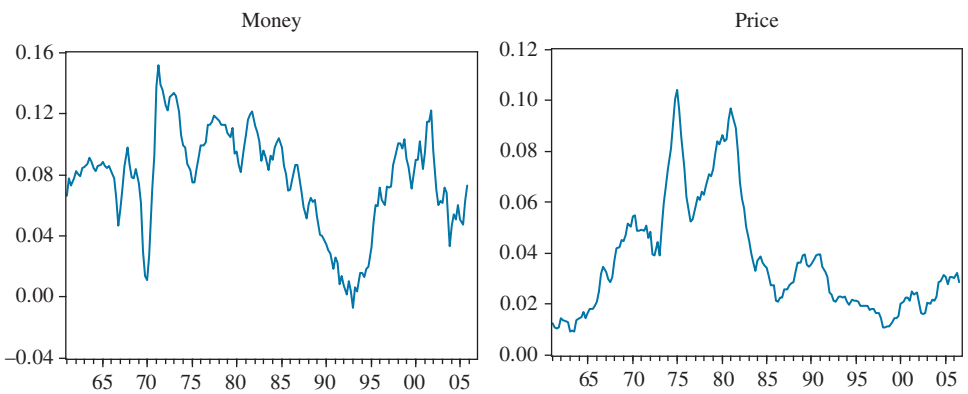


FIGURE 13.6 **Percentage changes in money and price.**

a. Identify the cointegrating relationship between P and M. Is the quantity theory of money supported?
b. Identify the error-correction coefficients. Is the system stable?
c. The above results were estimated using a system approach. Compute the cointegrating residuals and confirm that the series is indeed an I(0) variable.
d. Estimate a VEC model using the cointegrating residuals. (Your results should be the same as above.)

13.11 Research into the Phillips curve is concerned with providing empirical evidence of a tradeoff between inflation and unemployment. Can an economy experience lower unemployment if it is prepared to accept higher inflation? Figure 13.7 plots the changes in a measure of the unemployment rate (DU) and the changes in a measure of inflation (DP) for the United States for the sample period 1970M07 to 2009M06. A VAR model was estimated as follows:

$$\begin{aligned} DU_t &= 0.180DU_{t-1} - 0.046DP_{t-1} \\ (t) &\quad (3.905) \qquad (-0.909) \end{aligned}$$

$$\begin{aligned} DP_t &= -0.098DU_{t-1} + 0.373DP_{t-1} \\ (t) &\quad (-2.522) \qquad (8.711) \end{aligned}$$

FIGURE 13.7 **Changes in the unemployment and inflation rates.**

a. Is there evidence of an inverse relationship between the change in the unemployment rate (DU) and the change in the inflation rate (DP)?
b. What is the response of DU at time $t+1$ following a unit shock to DU at time t?
c. What is the response of DP at time $t+1$ following a unit shock to DU at time t?
d. What is the response of DU at time $t+2$?
e. What is the response of DP at time $t+2$?

13.12 Figure 13.8 shows the time series for two exchange rates—the *EURO* per $US and the *STERLING* per $US (data file *sterling*). Both the levels and the changes in the data are shown.

a. Which set of data would you consider using to estimate a VEC model, and which set to estimate a VAR? Why?
b. Apply the two-step approach suggested in this chapter to estimate a VEC model.
c. Estimate a VAR model paying attention to the order of the lag.

13.13 Financial analysts often debate the role of dividends (DV) in the determination of share prices (SP). Figure 13.9 shows plots of the rate of change in DV and SP computed as

$$DV_t = 100\ln\left(DN_t/DN_{t-1}\right), \quad SP_t = 100\ln\left(PN_t/PN_{t-1}\right)$$

where PN is the Standard and Poor Composite Price Index; DN is the nominal dividend per share (source: Prescott, E. C. and Mehra, R. "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, 15 March, 1985, pp. 145–161). The data are annual observations over the period 1889–1979.

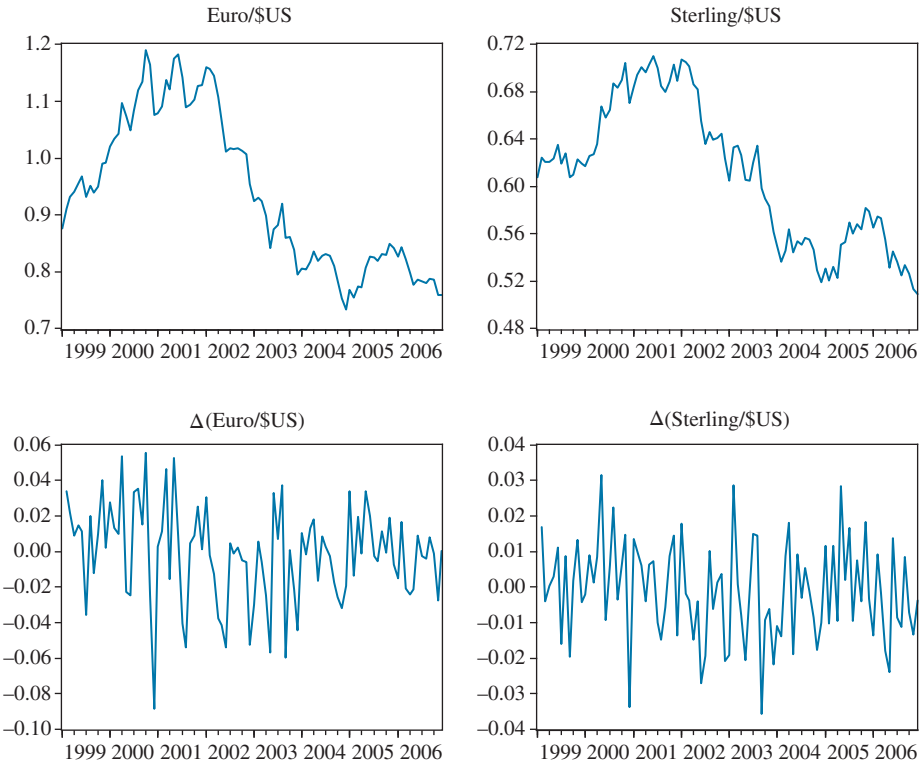


FIGURE 13.8 **Exchange rates.**

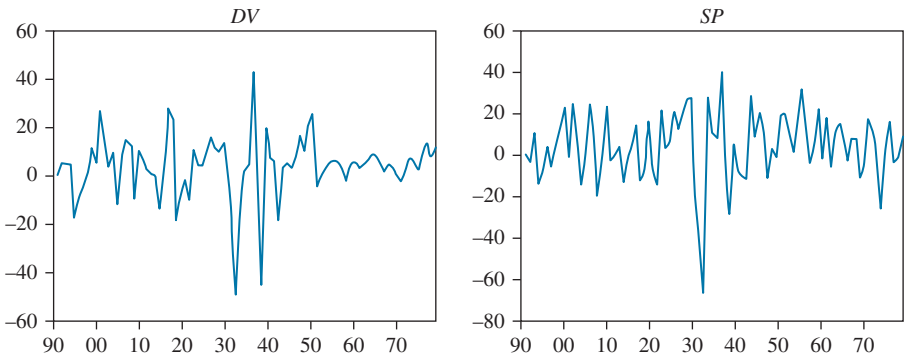


FIGURE 13.9 **Change in dividends (*DV*) and share price (*SP*).**

The data file is called *equity*. Estimate a first-order VAR for *SP* and *DV* by applying least squares to each equation:

$$SP_t = \beta_{10} + \beta_{11}SP_{t-1} + \beta_{12}DV_{t-1} + v_t^s$$
$$DV_t = \beta_{20} + \beta_{21}SP_{t-1} + \beta_{22}DV_{t-1} + v_t^d$$

Estimate an ARDL for each equation:

$$SP_t = \alpha_{10} + \alpha_{11}SP_{t-1} + \alpha_{12}DV_{t-1} + \alpha_{13}DV_t + e_t^s$$
$$DV_t = \alpha_{20} + \alpha_{21}SP_{t-1} + \alpha_{22}DV_{t-1} + \alpha_{23}SP_t + e_t^d$$

Compare the two sets of results and note the importance of the contemporaneous endogenous variable (*SP*, *DV*) in each equation.

a. Explain why least squares estimation of the VAR model with lagged variables on the right-hand side yields consistent estimates.
b. Explain why least squares estimation of the model with lagged and contemporaneous variables on the right-hand side yields inconsistent estimates. (You might like to refer to the material in Chapter 11.)
c. What do you infer about the role of dividends in the determination of share prices?

13.14 The file *gfc* contains data about economic activity in two major economies: the United States and the Euro Area (the group of countries in Europe where the Euro currency is the legal tender). Specifically, the data are the logs of their GDP, standardized so that the value of GDP is equal to 100 in 2000. The levels and the change in economic activity are shown in Figure 13.10(a) and (b). The sample period is from 1995Q1 to 2009Q4 and includes the global financial crisis that began in September 2007.

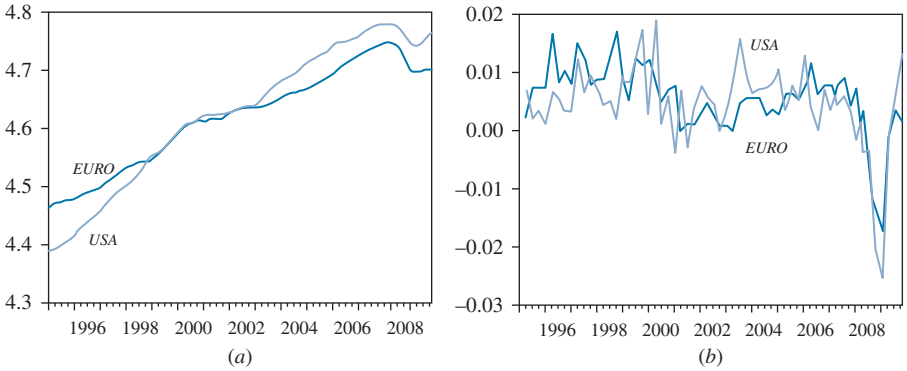


FIGURE 13.10 **Logs of GDP (*a*) and change in logs of GDP (*b*).**

a. Based on a visual inspection of the data, what would you infer about the interactions between the GDPs in the two economies?
b. Do the economies have a long-run relationship? Specify the econometric model and estimate the model. Plot the residuals and comment on their properties.
c. Do the economies have a short-run relationship? Specify the econometric model and estimate the model. Plot the residuals and comment on their properties.

13.15 The file *precious* contains monthly data on the prices of gold and silver (in logs) for the period 1970M1 to 2014M2.

a. Plot the two series and comment on the graph. Do the two prices appear to be moving together?
b. Use a series of hypothesis tests to decide on predictive models for the price of silver and the price of gold.

Appendix 13A The Identification Problem[3]

A bivariate dynamic system with contemporaneous interactions (also known as a structural model) is written as

$$y_t + \beta_1 x_t = \alpha_1 y_{t-1} + \alpha_2 x_{t-1} + e_t^y$$
$$x_t + \beta_2 y_t = \alpha_3 y_{t-1} + \alpha_4 x_{t-1} + e_t^x$$

[3]This appendix requires a basic understanding of matrix notation.

which can be more conveniently expressed in matrix form as

$$\begin{bmatrix} 1 & \beta_1 \\ \beta_2 & 1 \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} e_t^y \\ e_t^x \end{bmatrix}$$

or rewritten in symbolic form as $BY_t = AY_{t-1} + E_t$, where

$$Y_t = \begin{bmatrix} y_t \\ x_t \end{bmatrix} \quad B = \begin{bmatrix} 1 & \beta_1 \\ \beta_2 & 1 \end{bmatrix} \quad A = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \quad E_t = \begin{bmatrix} e_t^y \\ e_t^x \end{bmatrix}$$

A VAR representation (also known as reduced-form model) is written as

$$y_t = \delta_1 y_{t-1} + \delta_2 x_{t-1} + v_t^y$$
$$x_t = \delta_3 y_{t-1} + \delta_4 x_{t-1} + v_t^x$$

or in matrix form as: $Y_t = CY_{t-1} + V_t$, where

$$C = \begin{bmatrix} \delta_1 & \delta_2 \\ \delta_3 & \delta_4 \end{bmatrix} \quad V_t = \begin{bmatrix} v_t^y \\ v_t^x \end{bmatrix}$$

Clearly, there is a relationship between (13.A.1) and (13.A.2): $C = B^{-1}A$ and $V_t = B^{-1}E_t$. The special case considered in the chapter assumes that there are no contemporaneous interactions $(\beta_1 = \beta_2 = 0)$, making B an identity matrix. There is no identification problem in this case because the VAR residuals can be unambiguously "identified" as shocks to y or as shocks to x: $v^y = e^y$, $v^x = e^x$. The generation and interpretation of the impulse responses and variance decompositions are unambiguous.

In general, however, B is not an identity matrix, making v^y and v^x weighted averages of e^y and e^x. In this general case, impulse responses and variance decompositions based on v^y and v^x are not meaningful or useful because we cannot be certain about the source of the shocks. A number of methods exist for "identifying" the structural model from its reduced form.

CHAPTER 14

Time-Varying Volatility and ARCH Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the difference between a constant and a time-varying variance of the error term.
2. Explain the term ''conditionally normal.''
3. Perform a test for ARCH effects.
4. Estimate an ARCH model.
5. Forecast volatility.
6. Explain the difference between ARCH and GARCH specifications.
7. Explain the distinctive features of a T-GARCH model and a GARCH-in-mean model.

KEYWORDS

ARCH
ARCH-in-mean
conditionally normal
GARCH
GARCH-in-mean
T-ARCH and T-GARCH
time-varying variance

In Chapter 12, our focus was on time-varying mean processes and macroeconomic time series. We were concerned with stationary and nonstationary variables, and, in particular, macroeconomic variables like gross domestic product (GDP), inflation, and interest rates. The nonstationary nature of the variables implied that they had **means that change over time**. In this chapter, we are concerned with stationary series, but with conditional variances that change over time. The model we focus on is called the autoregressive conditional heteroskedastic (**ARCH**) model.

Nobel Prize winner Robert Engle's original work on ARCH was concerned with the volatility of inflation. However, it was applications of the ARCH model to financial time series that established and consolidated the significance of his contribution. For this reason, the examples used in this chapter will be based on financial time series. As we will see, financial time series have characteristics that are well represented by models with dynamic variances. The particular aims of this chapter are to discuss the modeling of dynamic variances using the ARCH class of models of volatility, the estimation of these models, and their use in forecasting.

14.1 The ARCH Model

ARCH stands for **a**uto**r**egressive **c**onditional **h**eteroskedasticity. We have covered the concepts of autoregressive and heteroskedastic errors in Chapters 9 and 8, respectively, so let us begin with a discussion of the concepts of conditional and unconditional means and variances of the error term.

Consider a model with an AR(1) error term

$$y_t = \phi + e_t \tag{14.1a}$$

$$e_t = \rho e_{t-1} + v_t, \qquad |\rho| < 1 \tag{14.1b}$$

$$v_t \sim N(0, \sigma_v^2) \tag{14.1c}$$

For convenience of exposition, first perform some successive substitution to obtain e_t as the sum of an infinite series of the error term v_t. To do this, note that if $e_t = \rho e_{t-1} + v_t$, then $e_{t-1} = \rho e_{t-2} + v_{t-1}$ and $e_{t-2} = \rho e_{t-3} + v_{t-2}$, and so on. Hence $e_t = v_t + \rho^2 v_{t-2} + \cdots + \rho^t e_0$ where the final term $\rho^t e_0$ is negligible.

The **unconditional mean** of the error is

$$E[e_t] = E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots] = 0$$

because $E[v_{t-j}] = 0$ for all j, whereas the **conditional mean** for the error, conditional on information prior to time t, is

$$E[e_t | I_{t-1}] = E[\rho e_{t-1} | I_{t-1}] + E[v_t] = \rho e_{t-1}$$

The information set at time $t-1$, I_{t-1}, includes knowing ρe_{t-1}. Put simply, "unconditional" describes the situation when you have no information, whereas conditional describes the situation when you have information, up to a certain point in time.

The **unconditional variance** of the error is

$$\begin{aligned} E[e_t - 0]^2 &= E[v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots]^2 \\ &= E[v_t^2 + \rho^2 v_{t-1}^2 + \rho^4 v_{t-2}^2 + \cdots] \\ &= \sigma_v{}^2[1 + \rho^2 + \rho^4 + \cdots] = \frac{\sigma_v{}^2}{1 - \rho^2} \end{aligned}$$

because $E[v_{t-j} v_{t-i}] = \sigma_v^2$ when $i = j$; $E[v_{t-j} v_{t-i}] = 0$ when $i \neq j$ and the sum of a geometric series $[1 + \rho^2 + \rho^4 + \cdots]$ is $1/(1 - \rho^2)$. The **conditional variance** for the error is

$$E\left[(e_t - \rho e_{t-1})^2 \middle| I_{t-1}\right] = E[v_t^2 | I_{t-1}] = \sigma_v^2$$

Now notice, for this model, that the conditional mean of the error varies over time, while the conditional variance does not. Suppose that instead of a conditional mean that changes over time, we have a conditional variance that changes over time. To introduce this modification, consider a variant of the above model

$$y_t = \beta_0 + e_t \tag{14.2a}$$

$$e_t | I_{t-1} \sim N(0, h_t) \tag{14.2b}$$

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2, \quad \alpha_0 > 0, \quad 0 \leq \alpha_1 < 1 \tag{14.2c}$$

Equations (14.2b and 14.2c) describe the ARCH class of models. The second equation (14.2b) says that the error term is **conditionally normal** $e_t | I_{t-1} \sim N(0, h_t)$ where I_{t-1} represents the information available at time $t-1$ with mean 0 and time-varying variance, denoted as h_t, following

popular terminology. The third equation (14.2c) models h_t as a function of a constant term and the lagged error squared e_{t-1}^2.

The name ARCH conveys the fact that we are working with time-varying variances (heteroskedasticity) that depend on (are conditional on) lagged effects (autocorrelation). This particular example is an ARCH(1) model since the time-varying variance h_t is a function of a constant term (α_0) plus a term lagged once, the square of the error in the previous period $(\alpha_1 e_{t-1}^2)$. The coefficients, α_0 and α_1, have to be positive to ensure a positive variance. The coefficient α_1 must be less than 1, or h_t will continue to increase over time, eventually exploding. Conditional normality means that the normal distribution is a function of known information at time $t-1$; i.e., when $t=2$, $e_2|I_1 \sim N(0, \alpha_0 + \alpha_1 e_1^2)$ and when $t=3$, $e_3|I_2 \sim N(0, \alpha_0 + \alpha_1 e_2^2)$, and so on. In this particular case, conditioning on I_{t-1} is equivalent to conditioning on the square of the error in the previous period e_{t-1}^2.

Note that while the conditional distribution of the error e_t is assumed to be normal, the unconditional distribution of the error e_t will not be normal. This is not an inconsequential consideration given that a lot of real-world data appear to be drawn from non-normal distributions.

We have noted that, conditional on e_{t-1}^2, the mean and variance of the error term e_t are zero and h_t, respectively. To find the mean and variance of the unconditional distribution of e_t, we note that, conditional on e_{t-1}^2, the standardized errors are standard normal, that is,

$$\left(\frac{e_t}{\sqrt{h_t}} \middle| I_{t-1} \right) = z_t \sim N(0,1)$$

Because this distribution does not depend on e_{t-1}^2, it follows that the unconditional distribution of $z_t = \left(e_t/\sqrt{h_t}\right)$ is also $N(0, 1)$, and that z_t and e_{t-1}^2 are independent. Thus, we can write

$$E(e_t) = E(z_t)E\left(\sqrt{\alpha_0 + \alpha_1 e_{t-1}^2}\right)$$

and

$$E(e_t^2) = E(z_t^2)E(\alpha_0 + \alpha_1 e_{t-1}^2) = \alpha_0 + \alpha_1 E(e_{t-1}^2)$$

From the first of these equations, we get $E(e_t) = 0$ because $E(z_t) = 0$. From the second of the equations, we get $\text{var}(e_t^2) = E(e_t^2) = \alpha_0/(1-\alpha_1)$ because $E(z_t^2) = 1$ and $E(e_t^2) = E(e_{t-1}^2)$.

The ARCH model has become a very important econometric model because it is able to capture stylized features of real-world volatility. Furthermore, in the context of the ARCH(1) model, knowing the squared error in the previous period e_{t-1}^2 improves our knowledge about the likely magnitude of the variance in period t. This is useful for situations when it is important to understand risk, as measured by the volatility of the variable.

14.2 Time-Varying Volatility

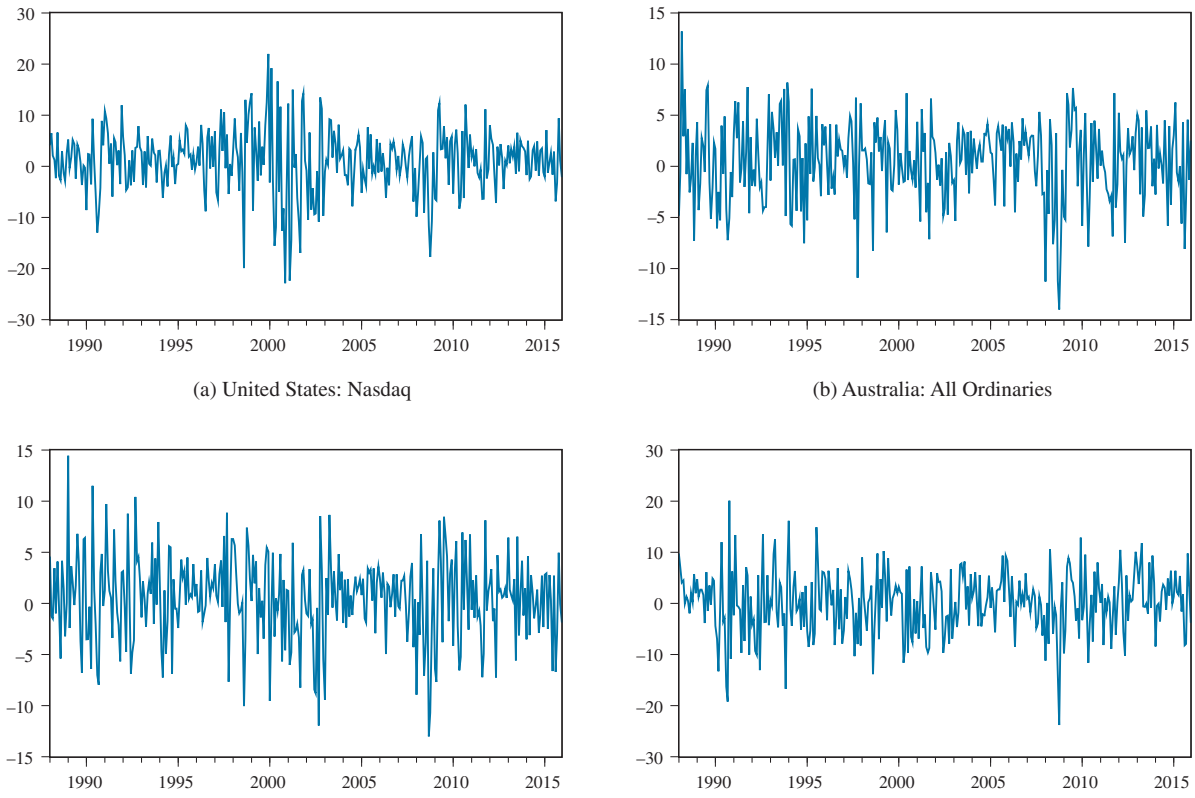
The ARCH model has become a popular one because its variance specification can capture commonly observed features of the time series of financial variables; in particular, it is useful for modeling **volatility** and especially changes in volatility over time. To appreciate what we mean by volatility and time-varying volatility, and how it relates to the ARCH model, let us look at some stylized facts about the behavior of financial variables—for example, the returns to stock price indices (also known as share price indices).

EXAMPLE 14.1 | Characteristics of Financial Variables

Figure 14.1 shows the time series of the monthly returns to a number of stock prices; namely, the U.S. Nasdaq, the Australian All Ordinaries, the Japanese Nikkei, and the UK FTSE over the period 1988M1 to 2015M12 (data file *returns5*). The values of these series change rapidly from period to period in an apparently unpredictable manner; we say the series are volatile. Furthermore, there are periods when large changes are followed by further large changes and periods when small changes are followed by further small changes. In this case the series are said to display time-varying volatility as well as "clustering" of changes.

Figure 14.2 shows the histograms of the returns. All returns display non-normal properties. We can see this more clearly if we draw normal distributions (using the respective sample means and sample variances) on top of these histograms. Note that there are more observations around the mean and in the tails. Distributions with these properties—more peaked around the mean and relatively fat tails—are said to be **leptokurtic**.

Note that the assumption that the conditional distribution for $(y_t|I_{t-1})$ is normal, an assumption that we made in (14.2b), does not necessarily imply that the unconditional distribution for y_t is normal. When we collect empirical observations on y_t into a histogram, we are constructing an estimate of the unconditional distribution for y_t. What we have observed is that the unconditional distribution for y_t is leptokurtic.



(a) United States: Nasdaq

(b) Australia: All Ordinaries

(c) United Kingdom: FTSE

(d) Japan: Nikkei

FIGURE 14.1 **Time series of returns to stock indices.**

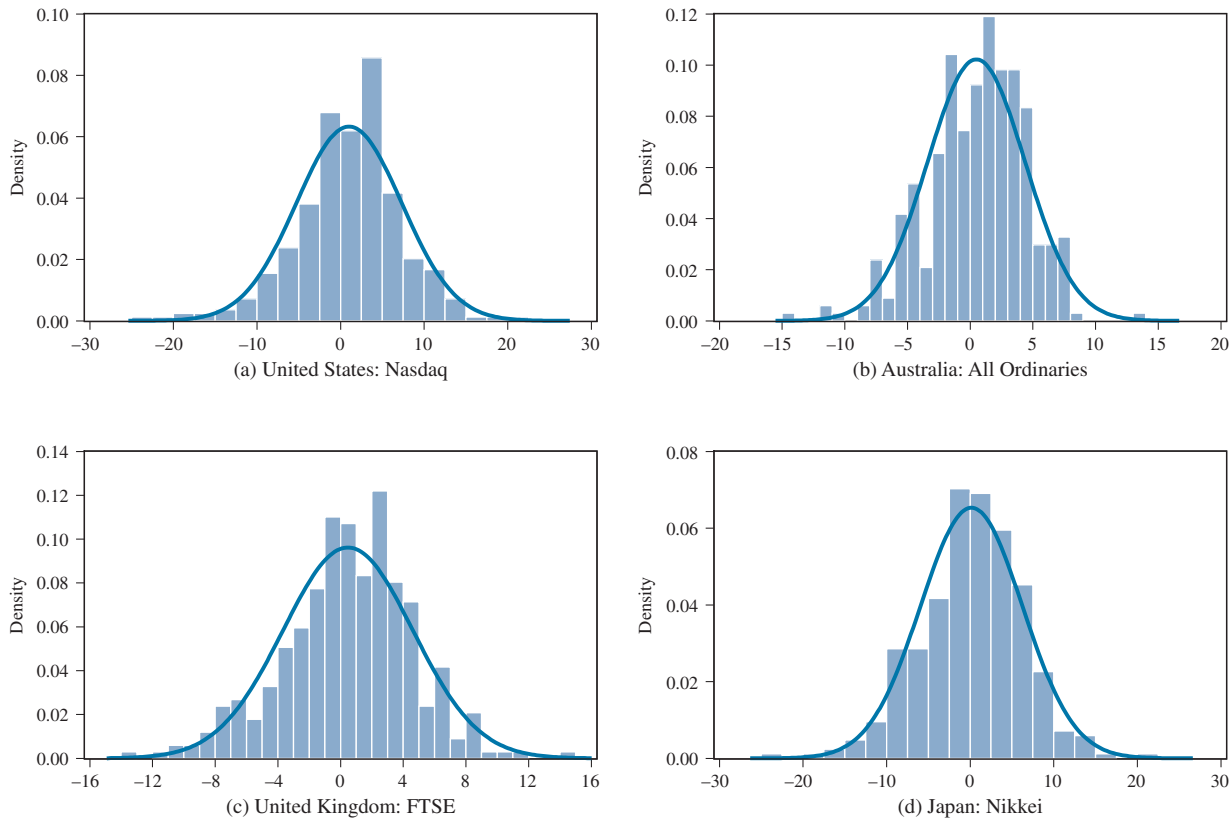


FIGURE 14.2 **Histograms of returns to stock indices.**

EXAMPLE 14.2 | Simulating Time-Varying Volatility

To illustrate how the ARCH model can be used to capture changing volatility and the leptokurtic nature of the distribution for y_t, we generate some simulated data for two models. In both cases we set $\beta_0 = 0$ so that $y_t = e_t$. The left panel in Figure 14.3 illustrates the case when $\alpha_0 = 1$, $\alpha_1 = 0$. These values imply $\text{var}(y_t|I_{t-1}) = h_t = 1$. This variance is constant, and not time varying, because $\alpha_1 = 0$. The right panel in Figure 14.3 illustrates the case when $\alpha_0 = 1$, $\alpha_1 = 0.8$, the case of a time-varying variance given by $\text{var}(y_t|I_{t-1}) = h_t = \alpha_0 + \alpha_1 e_{t-1}^2 = 1 + 0.8e_{t-1}^2$. Note that relative to the series in the left panel, volatility in the right panel is not constant; rather, it changes over time and it clusters—there are periods of small changes (e.g., around observation 100) and periods of big changes (around observation 175).

In Figure 14.4, we present histograms of y_t for the two cases. The top panel is the histogram for the constant variance case where $(y_t|I_{t-1})$ and y_t have the same distribution, namely the noise process $y_t \sim N(0, 1)$ because $h_t = 1$. The bottom panel is the histogram for the time-varying variance case. We know that the conditional distribution for $(y_t|I_{t-1})$

is $N(0, h_t)$. But what about the unconditional distribution for y_t? Again, we can check for normality by superimposing a normal distribution on top of the histogram. In this case, to allow for a meaningful comparison with the histogram in the top panel, we plot the standardized observations of y_t. That is, for each observation we subtract the sample mean and divide by the sample standard deviation. This transformation ensures that the distribution will have a zero mean and variance one, but it preserves the shape of the distribution. Comparing the two panels, we note that the second distribution has higher frequencies around the mean (zero) and higher frequencies in the tails (outside ± 3). This feature of time series with ARCH errors—the unconditional distribution of y_t is non-normal—is consistent with what we observed in the stock return series.

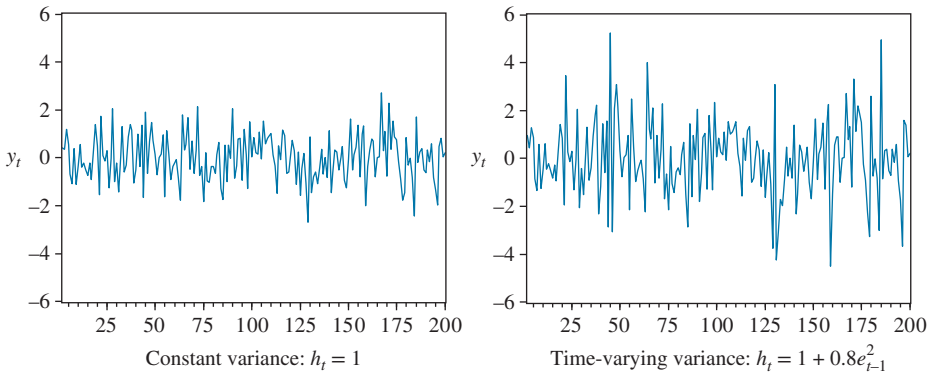


FIGURE 14.3 **Simulated examples of constant and time-varying variances.**

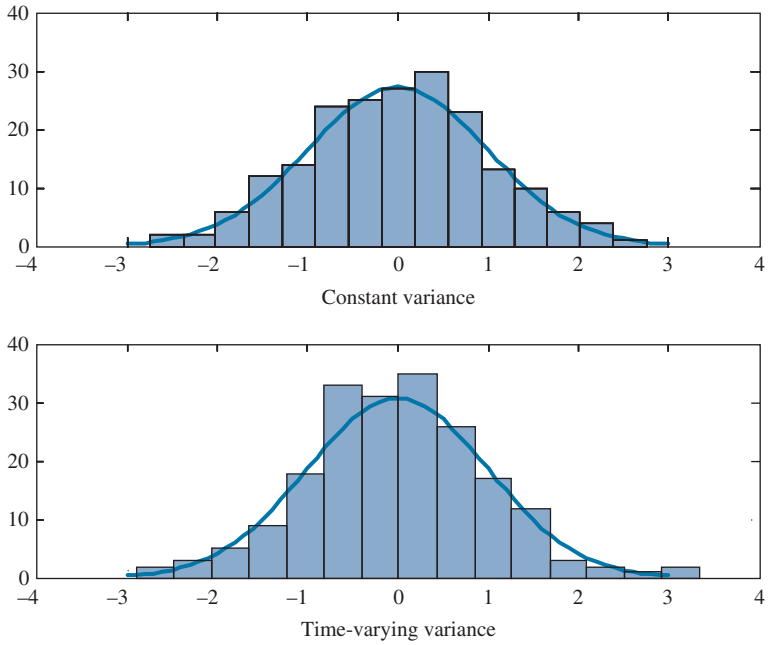


FIGURE 14.4 **Frequency distributions of the simulated models.**

Thus, the ARCH model is intuitively appealing because it seems sensible to explain volatility as a function of the errors e_t. These errors are often called "shocks" or "news" by financial analysts. They represent the unexpected! According to the ARCH model, the larger the shock, the greater the volatility in the series. In addition, this model captures volatility clustering, as big changes in e_t are fed into further big changes in h_t via the lagged effect e_{t-1}. The simulations show how well the ARCH model mimics the behavior of financial time series shown in Figure 14.1, including their non-normal distributions.

14.3 Testing, Estimating, and Forecasting

A Lagrange multiplier (LM) test is often used to test for the presence of ARCH effects. To perform this test, first estimate the **mean equation**, which can be a regression of the variable on a constant (like 14.1) or may include other variables. Then save the estimated residuals $\hat{e}_t$ and obtain their squares $\hat{e}_t^2$. To test for first-order ARCH, regress $\hat{e}_t^2$ on the squared residuals lagged $\hat{e}_{t-1}^2$,

$$\hat{e}_t^2 = \gamma_0 + \gamma_1 \hat{e}_{t-1}^2 + v_t \tag{14.3}$$

where v_t is a random term. The null and alternative hypotheses are

$$H_0 : \gamma_1 = 0 \qquad H_1 : \gamma_1 \neq 0$$

If there are no ARCH effects, then $\gamma_1 = 0$ and the fit of (14.3) will be poor, and the equation R^2 will be low. If there are ARCH effects, we expect the magnitude of $\hat{e}_t^2$ to depend on its lagged values, and the R^2 will be relatively high. The LM test statistic is $(T-q)R^2$ where T is the sample size, q is the number of $\hat{e}_{t-j}^2$ terms on the right-hand side of (14.3), and R^2 is the coefficient of determination. If the null hypothesis is true, then the test statistic $(T-q)R^2$ is distributed (in large samples) as $\chi^2_{(q)}$, where q is the order of lag, and $T-q$ is the number of complete observations; in this case, $q=1$. If $(T-q)R^2 \geq \chi^2_{(1-\alpha,\,q)}$, then we reject the null hypothesis that $\gamma_1 = 0$ and conclude that ARCH effects are present.

EXAMPLE 14.3 | Testing for ARCH in BrightenYourDay (BYD) Lighting

To illustrate the test, consider the returns from buying shares in the hypothetical company BYD Lighting. The time series and histogram of the returns are shown in Figure 14.5 (data file *byd*). The time series shows evidence of time-varying volatility and clustering, and the unconditional distribution is non-normal.

To perform the test for ARCH effects, first estimate a mean equation that in this example is $r_t = \beta_0 + e_t$, where r_t is the monthly return on shares of BYD. Second, retrieve the estimated residuals. Third, estimate (14.3). The results for the ARCH test are

$$\hat{e}_t^2 = 0.908 + 0.353\hat{e}_{t-1}^2 \qquad R^2 = 0.124$$
$$(t) \qquad\quad (8.409)$$

The t-statistic suggests a significant first-order coefficient. The sample size is 500, giving an LM test value of $(T-q)R^2 = 61.876$. Comparing the computed test value to the 5% critical value of a $\chi^2_{(1)}$ distribution $\left(\chi^2_{(0.95,1)} = 3.841\right)$ leads to the rejection of the null hypothesis. In other words, the residuals show the presence of ARCH(1) effects.

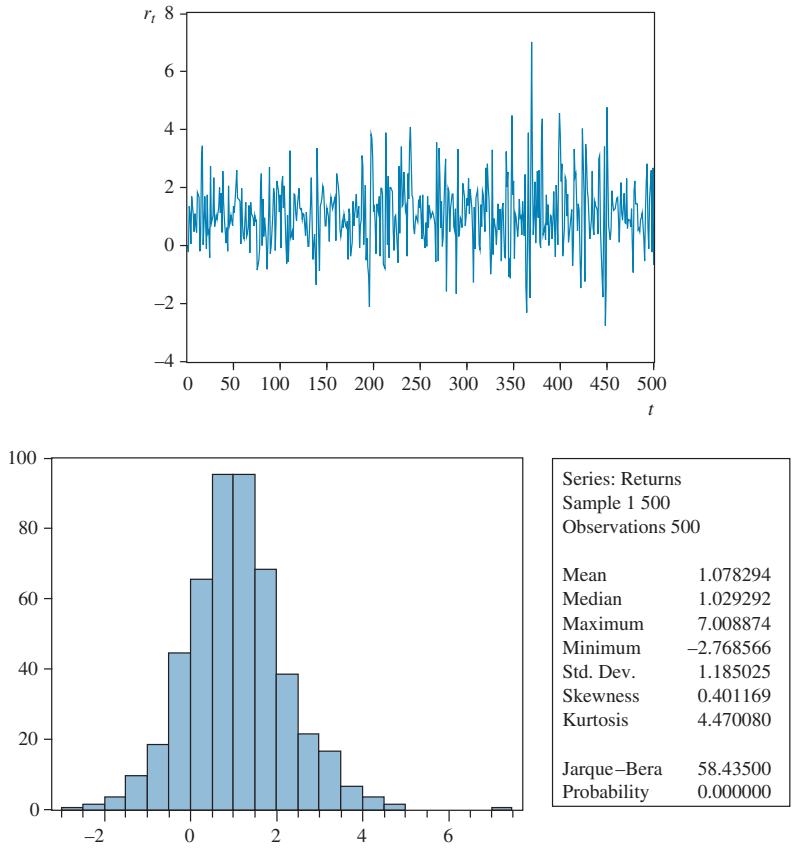


FIGURE 14.5 **Time series and histogram of returns for BYD lighting.**

EXAMPLE 14.4 | ARCH Model Estimates for BrightenYourDay (BYD) Lighting

ARCH models are estimated by the maximum likelihood method. Estimation details are beyond the scope of this book, but the maximum likelihood method (see Appendix C.8) is programmed in most econometric software.

Equation (14.4) shows the results from estimating an ARCH(1) model applied to the monthly returns from buying shares in BrightenYourDay Lighting. The estimated mean of the series is described in (14.4a), while the estimated variance is given in (14.4b).

$$\hat{r}_t = \hat{\beta}_0 = 1.063 \qquad (14.4a)$$

$$\hat{h}_t = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{e}^2_{t-1} = 0.642 + 0.569\hat{e}^2_{t-1} \qquad (14.4b)$$
$$(t) \qquad (5.536)$$

The t-statistic of the first-order coefficient (5.536) suggests a significant ARCH(1) coefficient. Recall that one of the requirements of the ARCH model is that $\alpha_0 > 0$ and $\alpha_1 > 0$, so that the implied variances are positive. Note that the estimated coefficients $\hat{\alpha}_0$ and $\hat{\alpha}_1$ satisfy this condition.

EXAMPLE 14.5 | Forecasting BrightenYourDay (BYD) Volatility

Once we have estimated the model, we can use it to forecast next period's return r_{t+1} and the conditional volatility h_{t+1}. When one invests in shares (or stocks), it is important to choose them not just on the basis of their mean returns, but also on the basis of their risk. Volatility gives us a measure of their risk.

For our case study of investing in BYD Lighting, the forecast return and volatility are

$$\hat{r}_{t+1} = \hat{\beta}_0 = 1.063 \tag{14.5a}$$

$$\hat{h}_{t+1} = \hat{\alpha}_0 + \hat{\alpha}_1\left(r_t - \hat{\beta}_0\right)^2 = 0.642 + 0.569\left(r_t - 1.063\right)^2 \tag{14.5b}$$

Equation (14.5a) gives the estimated return that—because it does not change over time—is both the conditional and unconditional mean return. The estimated error in period t, given by $\hat{e}_t = r_t - \hat{r}_t$, can then be used to obtain the estimated conditional variance (14.5b). The time series of the conditional variance does change over time and is shown in Figure 14.6. Note how the conditional variance around observation 370 coincides with the period of large changes in returns as shown in Figure 14.5.

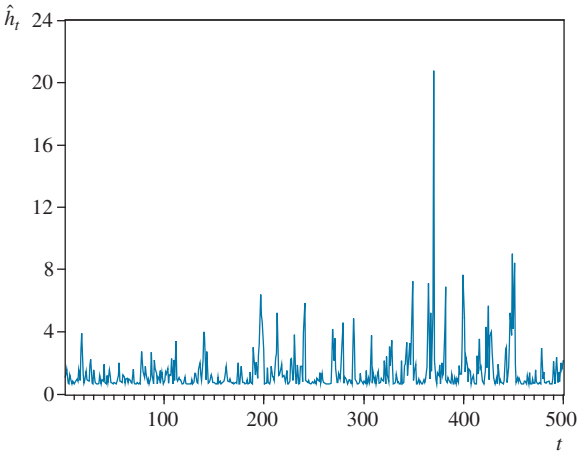


FIGURE 14.6 **Plot of conditional variance.**

14.4 Extensions

The ARCH(1) model can be extended in a number of ways. One obvious extension is to allow for more lags. In general, an ARCH(q) model that includes lags $\hat{e}^2_{t-1}, \ldots, \hat{e}^2_{t-q}$ has a conditional variance function that is given by

$$h_t = \alpha_0 + \alpha_1 e^2_{t-1} + \alpha_2 e^2_{t-2} \cdots + \alpha_q e^2_{t-q} \tag{14.6}$$

In this case the variance or volatility in a given period depends on the magnitudes of the squared errors in the past q periods. Testing, estimating, and forecasting, are natural extensions of the case with one lag.

14.4.1 The GARCH Model—Generalized ARCH

One of the shortcomings of an ARCH(q) model is that there are $q + 1$ parameters to estimate. If q is a large number, we may lose accuracy in the estimation. The generalized ARCH model, or **GARCH**, is an alternative way to capture long lagged effects with fewer parameters. It is a special generalization of the ARCH model and it can be derived as follows. First, consider (14.6) but write it as

$$h_t = \alpha_0 + \alpha_1 e^2_{t-1} + \beta_1\alpha_1 e^2_{t-2} + \beta_1^2\alpha_1 e^2_{t-3} + \cdots$$

In other words, we have imposed a geometric lag structure on the lagged coefficients of the form $\alpha_s = \alpha_1\beta_1^{s-1}$. Next, add and subtract $\beta_1\alpha_0$ and rearrange terms as follows:

$$h_t = \left(\alpha_0 - \beta_1\alpha_0\right) + \alpha_1 e^2_{t-1} + \beta_1\left(\alpha_0 + \alpha_1 e^2_{t-2} + \beta_1\alpha_1 e^2_{t-3} + \cdots\right)$$

Then, since $h_{t-1} = \alpha_0 + \alpha_1 e^2_{t-2} + \beta_1\alpha_1 e^2_{t-3} + \beta_1^2\alpha_1 e^2_{t-4} + \cdots$, we may simplify to

$$h_t = \delta + \alpha_1 e^2_{t-1} + \beta_1 h_{t-1} \tag{14.7}$$

where $\delta = \left(\alpha_0 - \beta_1\alpha_0\right)$. This generalized ARCH model is denoted as GARCH(1, 1). It can be viewed as a special case of the more general GARCH (p, q) model, where p is the number of lagged h terms and q is the number of lagged e^2 terms. We also note that we need $\alpha_1 + \beta_1 < 1$ for stationarity; if $\alpha_1 + \beta_1 \geq 1$ we have a so-called "integrated GARCH" process, or IGARCH.

The GARCH(1, 1) model is a very popular specification because it fits many data series well. It tells us that the volatility changes with lagged shocks $\left(e^2_{t-1}\right)$ but there is also momentum in

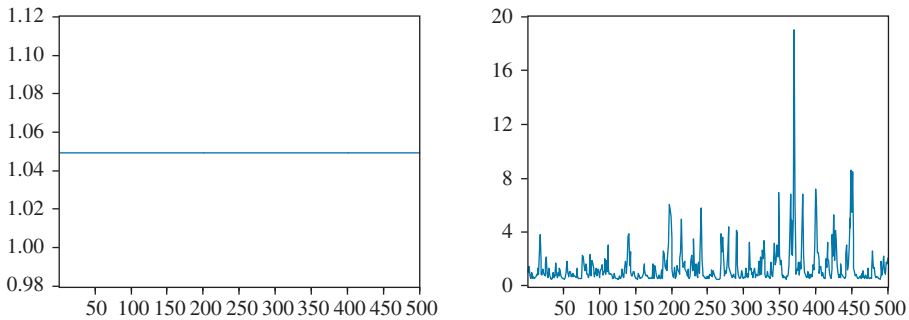
the system working via h_{t-1}. One reason why this model is so popular is that it can capture long lags in the shocks with only a few parameters. A GARCH(1, 1) model with three parameters $(\delta, \alpha_1, \text{ and } \beta_1)$ can capture similar effects to an ARCH(q) model requiring the estimation of $(q+1)$ parameters, where q is large, say $q \geq 6$.

EXAMPLE 14.6 | A GARCH Model for BrightenYourDay

To illustrate the GARCH(1, 1) specification, consider again the returns to our shares in BYD Lighting, which we re-estimate (by maximum likelihood) under the new model. The results are

$$\hat{r}_t = 1.049$$
$$\hat{h}_t = 0.401 + 0.492\,\hat{e}^2_{t-1} + 0.238\,\hat{h}_{t-1}$$
$$(t) \qquad (4.834) \qquad (2.136)$$

The significance of the coefficient in front of $\hat{h}_{t-1}$ suggests that the GARCH(1, 1) model is better than the ARCH(1) results shown in (14.4). Plots of the mean equation and the time-varying variance are shown in Figures 14.7(a) and (b), respectively.



(a) GARCH(1, 1): $E(r_t) = 1.049$

(b) GARCH(1, 1): $h_t = 0.401 + 0.492e^2_{t-1} + 0.238h_{t-1}$

FIGURE 14.7 **Estimated mean and variance of GARCH model.**

14.4.2 Allowing for an Asymmetric Effect

A standard ARCH model treats bad "news" $(\text{negative } e_{t-1} < 0)$ and good "news" $(\text{positive } e_{t-1} > 0)$ symmetrically, that is, the effect on the volatility h_t is the same $(\alpha_1 e^2_{t-1})$. However, the effects of good and bad news may have asymmetric effects on volatility. In general, when negative news hits a financial market, asset prices tend to enter a turbulent phase and volatility increases, but with positive news volatility tends to be small and the market enters a period of tranquility.

The threshold ARCH model, or **T-ARCH**, is one example where positive and negative news are treated asymmetrically. In the T-GARCH version of the model, the specification of the conditional variance is

$$h_t = \delta + \alpha_1 e^2_{t-1} + \gamma d_{t-1} e^2_{t-1} + \beta_1 h_{t-1}$$
$$d_t = \begin{cases} 1 & e_t < 0 \quad \text{(bad news)} \\ 0 & e_t \geq 0 \quad \text{(good news)} \end{cases} \tag{14.8}$$

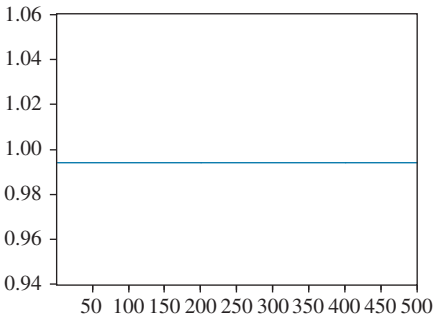
where γ is known as the asymmetry or leverage term. When $\gamma = 0$, the model collapses to the standard GARCH form. Otherwise, when the shock is positive (i.e., good news) the effect on volatility is α_1, but when the news is negative (i.e., bad news) the effect on volatility is $\alpha_1 + \gamma$. Hence, if γ is significant and positive, negative shocks have a larger effect on h_t than positive shocks.

EXAMPLE 14.7 | A T-GARCH Model for BYD

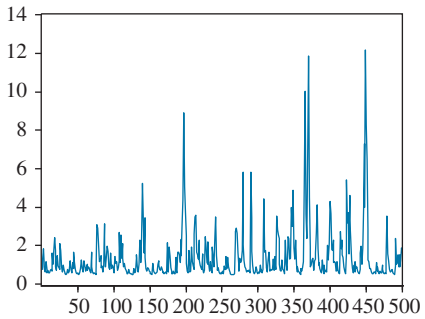
The returns to our shares in BYD Lighting were re-estimated with a T-GARCH(1,1) specification:

$$\hat{r}_t = 0.994$$
$$\hat{h}_t = 0.356 + 0.263\hat{e}^2_{t-1} + 0.492d_{t-1}\hat{e}^2_{t-1} + 0.287\hat{h}_{t-1}$$
$$(t) \qquad (3.267) \qquad (2.405) \qquad (2.488)$$

These results show that when the market observes good news (positive e_t), the contribution of e_t^2 to volatility h_{t+1} is by a factor 0.263, whereas when the market observes bad news (negative e_t), the contribution of e_t^2 to volatility h_{t+1} is by a factor $(0.263 + 0.492)$. Overall, negative shocks create greater volatility in financial markets. The mean and variance are displayed in Figures 14.8(a) and (b). Note that, relative to Figure 14.7(b), the T-GARCH model has highlighted the period around observation 200 as another period of turbulence.



(a) T-GARCH(1, 1): $E(r_t) = 0.994$



(b) T-GARCH(1, 1):
$h_t = 0.356 + (0.263 + 0.492d_{t-1})e^2_{t-1} + 0.287h_{t-1}$

FIGURE 14.8 **Estimated mean and variance of T-GARCH model.**

14.4.3 GARCH-in-Mean and Time-Varying Risk Premium

Another popular extension of the GARCH model is the **GARCH-in-mean** model. The positive relationship between risk (often measured by volatility) and return is one of the basic tenets of financial economics. As risk increases, so does the mean return. Intuitively, the return to risky assets tends to be higher than the return to safe assets (low variation in returns) to compensate an investor for taking on the risk of buying the volatile share. However, while we have estimated the mean equation to model returns, and have estimated a GARCH model to capture time-varying volatility, we have not used the risk to explain returns. This is the aim of the GARCH-in-mean models.

The equations of a GARCH-in-mean model are shown below:

$$y_t = \beta_0 + \theta h_t + e_t \tag{14.9a}$$
$$e_t | I_{t-1} \sim N(0, h_t) \tag{14.9b}$$
$$h_t = \delta + \alpha_1 e^2_{t-1} + \beta_1 h_{t-1}, \quad \delta > 0, \quad 0 \le \alpha_1 < 1, \quad 0 \le \beta_1 < 1 \tag{14.9c}$$

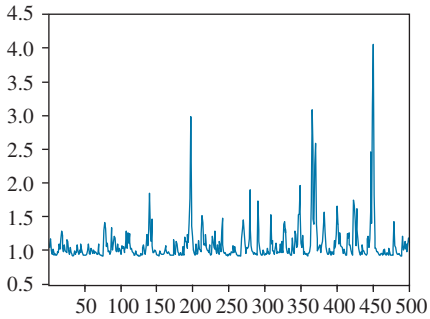
The first equation is the mean equation; it now shows the effect of the conditional variance on the dependent variable. In particular, note that the model postulates that the conditional variance h_t affects y_t by a factor θ. The other two equations are as before.

EXAMPLE 14.8 | A GARCH-in-Mean Model for BYD

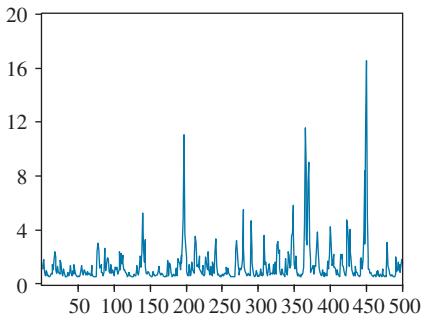
The returns to shares in BYD Lighting were reestimated as a GARCH-in-mean model:

$$\hat{r}_t = 0.818 + 0.196h_t$$
$$(t) \qquad (2.915)$$
$$\hat{h}_t = 0.370 + 0.295\hat{e}^2_{t-1} + 0.321d_{t-1}\hat{e}^2_{t-1} + 0.278\hat{h}_{t-1}$$
$$(t) \qquad (3.426) \qquad (1.979) \qquad (2.678)$$

The results show that as volatility increases, the returns correspondingly increase by a factor of 0.196. In other words, this result supports the usual view in financial markets—high risk, high return. The GARCH-in-mean model is shown in Figures 14.9(a) and (b). Note that the expected mean return is no longer a constant value, but rather has high values (e.g., around observation 200) that coincide with higher conditional variances.



(a) GARCH-in-mean: $E(r_t) = 0.818 + 0.196h_t$



(b) GARCH-in-mean: $h_t = 0.370 + (0.295 + 0.321d_{t-1})e^2_{t-1} + 0.278h_{t-1}$

FIGURE 14.9 **Estimated mean and variance of GARCH-in-mean model.**

One last point before we leave this section. The first equation of the GARCH-in-mean model is sometimes written as a function of the time-varying standard deviation $\sqrt{h_t}$, that is, $y_t = \beta_0 + \theta\sqrt{h_t} + e_t$. This is because both measures—variance and standard deviation—are used by financial analysts to measure risk. There are no hard-and-fast rules about which measure to use. Exercise 14.8 illustrates the case when we use $\sqrt{h_t}$. A standard t-test of significance is often used to decide which is the more suitable measure.

14.4.4 Other Developments

The GARCH, T-GARCH, and GARCH-in-mean models are three important extensions of the original ARCH concept developed by Engle in 1982. There have also been numerous other variations, developed to handle complexities noted in the data, especially in high frequency financial data. One variation, exponential GARCH (EGARCH), has stood the test of time. This model is

$$\ln(h_t) = \delta + \beta_1 \ln(h_{t-1}) + \alpha\left|\frac{e_{t-1}}{\sqrt{h_{t-1}}}\right| + \gamma\left(\frac{e_{t-1}}{\sqrt{h_{t-1}}}\right)$$

where $\left(e_{t-1}/\sqrt{h_{t-1}}\right)$ are the standardized residuals. The model uses a log specification, which ensures the estimated variance remains positive. It also includes two standardized residual terms, with one of them in absolute form to facilitate the testing of the leverage effect. The **leverage effect** refers to the generally observed negative correlation between an asset return and its volatility changes. One potential explanation for this observation is that bad news has a bigger effect on

variance than good news. If $\gamma \neq 0$, the effects of good/bad news are asymmetric; if $\gamma < 0$, negative shocks have larger effects.

Another significant development is to allow the conditional distribution of the error term to be non-normal. Because empirical distributions of financial returns generally exhibit fat tails and clustering around zero, the *t*-distribution has become a popular alternative to the assumption of normality. Also, regressors have been introduced in the variance equation to allow volatility to depend on exogenous or predetermined variables. Shift (dummy) variables are especially popular and have been used to allow for changes in political regimes.

14.5 Exercises

14.5.1 Problems

14.1 The ARCH model is sometimes presented in the following multiplicative form:

$$y_t = \beta_0 + e_t$$
$$e_t = z_t\sqrt{h_t},\ z_t \sim N(0,1)$$
$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2,\ \alpha_0 > 0,\ 0 \leq \alpha_1 < 1$$

This form describes the distribution of the standardized residuals $e_t/\sqrt{h_t}$ as standard normal z_t. However, the properties of e_t are not altered.

a. Show that the conditional mean $E(e_t|I_{t-1}) = 0$.
b. Show that the conditional variance $E(e_t^2|I_{t-1}) = h_t$.
c. Show that $e_t|I_{t-1} \sim N(0, h_t)$.

14.2 The equations of an **ARCH-in-mean** model are shown below:

$$y_t = \beta_0 + \theta h_t + e_t$$
$$e_t|I_{t-1} \sim N(0, h_t)$$
$$h_t = \delta + \alpha_1 e_{t-1}^2,\ \delta > 0,\ 0 \leq \alpha_1 < 1$$

Let y_t represent the return from a financial asset and let e_t represent "news" in the financial market. Now use the third equation to substitute out h_t in the first equation, to express the return as

$$y_t = \beta_0 + \theta(\delta + \alpha_1 e_{t-1}^2) + e_t$$

a. If θ is zero, what is $E_t(y_{t+1})$, the conditional mean of y_{t+1}? In other words, what do you expect next period's return to be, given information today?
b. If θ is not zero, what is $E_t(y_{t+1})$? What extra information have you used here to forecast the return?

14.3 Consider the following T-ARCH model:

$$h_t = \delta + \alpha_1 e_{t-1}^2 + \gamma d_{t-1} e_{t-1}^2$$
$$d_t = \begin{cases} 1 & e_t < 0 \quad \text{(bad news)} \\ 0 & e_t \geq 0 \quad \text{(good news)} \end{cases}$$

a. If γ is zero, what are the values of h_t when $e_{t-1} = -1$, when $e_{t-1} = 0$, and when $e_{t-1} = 1$?
b. If γ is not zero, what are the values of h_t when $e_{t-1} = -1$, when $e_{t-1} = 0$, and when $e_{t-1} = 1$? What is the key difference between the case $\gamma = 0$ and $\gamma \neq 0$?

14.4 The GARCH(1, 1) model shown below can also be reexpressed as an ARCH(q) model, where q is a large number (in fact, infinity). Derive the ARCH form of a GARCH model using the method of recursive substitution.

$$h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$$

14.5 **a.** Let $I_{t-1} = \{e_{t-1}, e_{t-2}, \ldots\}$. Use the law of iterated iterations to show that $E(e_t|I_{t-1}) = 0$ implies $E(e_t) = 0$.

b. Consider the variance model $h_t = E(e_t^2|I_{t-1}) = \alpha_0 + \alpha_1 e_{t-1}^2$. Use the law of iterated iterations to show that, for $0 < \alpha_1 < 1$, $E(e_t^2) = \alpha_0/(1-\alpha_1)$.

c. Consider the variance model $h_t = E(e_t^2|I_{t-1}) = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$. Use the law of iterated iterations to show that for $0 < \alpha_1 + \beta_1 < 1$, $E(e_t^2) = \delta/(1-\alpha_1-\beta_1)$.

14.6 The estimates for the five models in Table 14.1 were obtained using monthly observations on returns to U.S. Nasdaq stock prices from 1985M1 to 2015M12. Use each of the models to estimate the mean and variance of returns for 2016M1.

TABLE 14.1 **Estimates from ARCH Models for U.S. Nasdaq Returns**

Mean function					
Constant	1.4567	1.1789	1.098	1.078	0.931
h_t					0.006
Variance function					
Constant	23.35	19.35	2.076	2.351	2.172
e_{t-1}^2	0.4694	0.3429	0.1329	0.124	0.136
e_{t-2}^2		0.1973			
h_{t-1}			0.8147	0.8006	0.8089
$d_{t-1}e_{t-1}^2$				0.0293	
End-of-sample estimates					
$\hat{e}_{2015M12}$	−3.4388	−3.1610	−3.0803	−3.0605	−3.0760
$\hat{e}_{2015M11}$	−0.3700	−0.0922	−0.0115	0.0083	−0.0296
$\hat{h}_{2015M12}$	23.42	32.64	27.10	27.39	27.27

14.5.2 Computer Exercises

14.7 The data file *share* contains time-series data on the Straits Times share price index of Singapore.

a. Compute the time series of returns using the formula $r_t = 100\ln(y_t/y_{t-1})$, where y_t is the share price index. Generate the correlogram of returns up to at least order 12, since the frequency of the data is monthly. Is there evidence of autocorrelation? If so, it indicates the presence of significant lagged mean effects.

b. Square the returns and generate the correlogram of squared returns. Is there evidence of significant lagged effects? If so, it indicates the presence of significant lagged variance effects.

14.8 The data file *euro* contains 204 monthly observations on the returns to the Euro share price index for the period 1988M1 to 2004M12. A plot of the returns data is shown in Figure 14.10(a), together with its histogram in Figure 14.10(b).

a. What do you notice about the volatility of returns? Identify the periods of big changes and the periods of small changes.

b. Is the distribution of returns normal? Is this the unconditional, or conditional, distribution?

c. Perform a LM test for the presence of first-order ARCH and check that you obtain the following results:

$$\hat{e}_t^2 = 20.509 + 0.237\hat{e}_{t-1}^2 \qquad (T-1)R^2 = 11.431$$
$$(t) \qquad\quad (3.463)$$

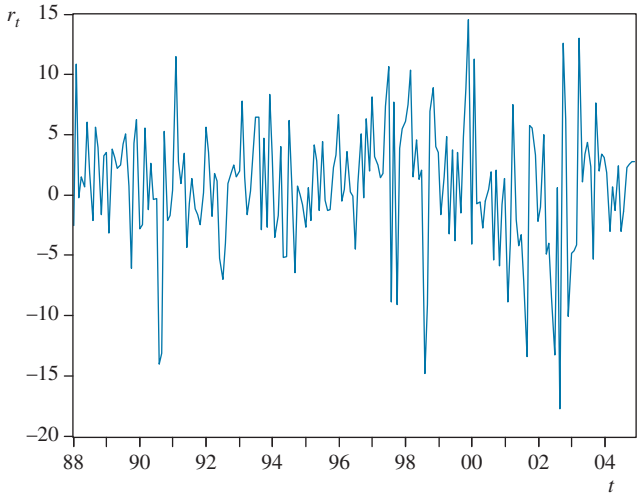
Is there evidence of ARCH effects?

d. Estimate an ARCH(1) model and check that you obtain the following results:

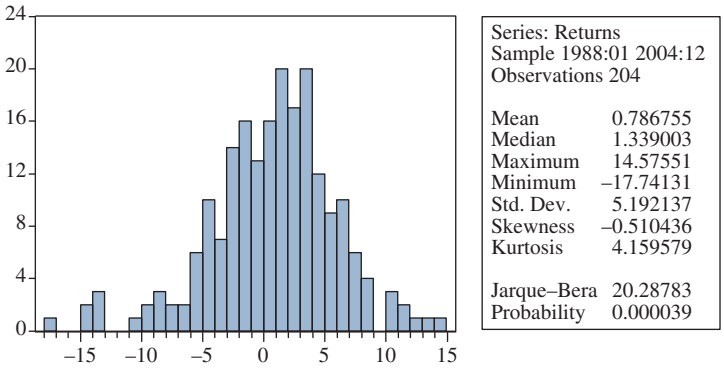
$$\hat{r}_t = 0.879 \qquad \hat{h}_t = 20.604 + 0.230\hat{e}^2_{t-1}$$
$$(t) \quad (2.383) \qquad\qquad (10.968) \quad (2.198)$$

Interpret the results.

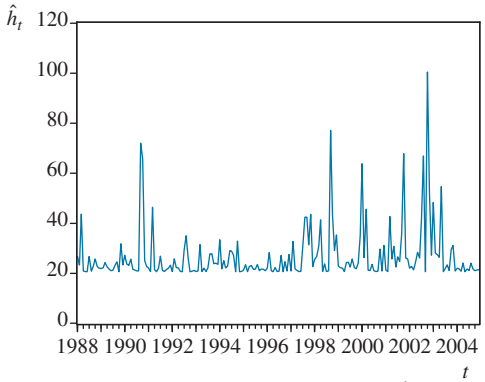
e. A plot of the conditional variance is shown in Figure 14.10(c). Do the periods of high and low conditional variance coincide with the periods of big and small changes in returns?



(a) Returns to Euro share price index



(b) Histogram of returns



(c) Estimated conditional variance $\hat{h}_t$

FIGURE 14.10 **Graphs for Exercise 14.8.**

14.9 Monthly changes in the \$US/\$AUS exchange rate S_t for the period 1985M7 to 2010M6 are stored in the file *exrate5*.

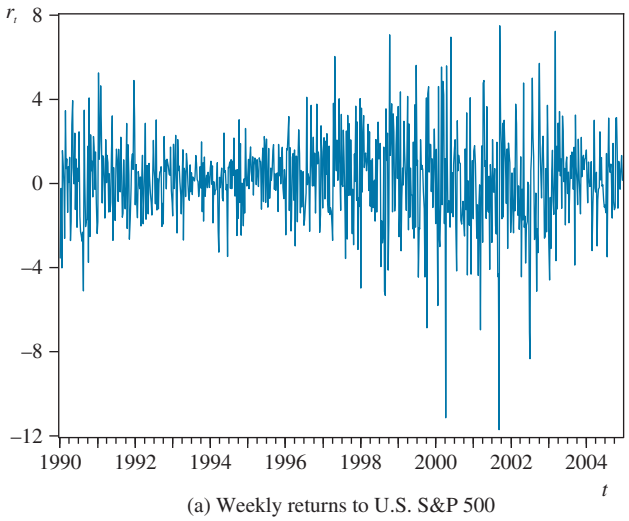
a. Plot the time series of the changes and their histogram. Are there periods of high volatility and periods of low volatility? Does the unconditional distribution of the changes appear to be normally distributed?

b. Estimate the GARCH(1, 1) model $S_t = \beta_0 + e_t, (e_t|I_{t-1}) \sim N(0, h_t)$ and $h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$. Comment on the results.

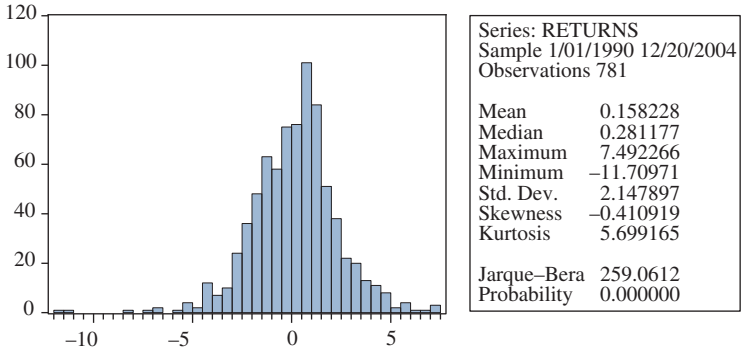
c. Estimate the conditional variance h_t for each observation and create the series $v_t = \hat{e}_t / \sqrt{\hat{h}_t}$ where $\hat{e}_t$ are the residuals $\hat{e}_t = S_t - \hat{\beta}_0$. Create a histogram for the v_t. Do they appear to be normally distributed?

d. Forecast the conditional mean and variance for 2010M7 and 2010M8.

14.10 Figure 14.11 shows the weekly returns to the U.S. S&P 500 for the sample period January 1990 to December 2004 (data file *sp*).



(a) Weekly returns to U.S. S&P 500



(b) Histogram of returns

FIGURE 14.11 **Graphs for Exercise 14.10.**

a. Estimate an ARCH(1) model and check that you obtain the following results:

$$\hat{r}_t = 0.197 \qquad \hat{h}_t = 3.442 + 0.253\hat{e}_{t-1}^2$$
$$(t) \quad (2.899) \qquad (22.436) \quad (5.850)$$

What is the value of the conditional variance when the last period's shock was positive, $e_{t-1} = +1$? What about when the last period's shock was negative, $e_{t-1} = -1$?

b. Estimate a T-ARCH model and check that you obtain the following results:

$$\hat{r}_t = 0.147 \qquad \hat{h}_t = 3.437 + (0.123 + 0.268d_{t-1})\hat{e}^2_{t-1}$$
$$(t) \quad (2.049) \qquad (22.963) \quad (2.330) \quad (2.944)$$

c. What is the value of the conditional variance when the last period's shock was positive, $e_{t-1} = +1$? When the last period's shock was negative, $e_{t-1} = -1$?

d. Is the asymmetric T-ARCH model better than the symmetric ARCH model in a financial econometric sense? [*Hint*: Look at the statistical tests for significance.] Is the asymmetric T-ARCH model better than the symmetric ARCH model in a financial economic sense? [*Hint*: Look at the implications of the results.]

14.11 Figure 14.12 shows the daily term premiums between a 180-day bank bill rate and a 90-day bank rate for the period July 1996 to December 1998 (data file *term*). Preliminary unit root tests confirm that the series may be treated as a stationary series, although the value of ρ, the autocorrelation coefficient, is quite high (about 0.9).

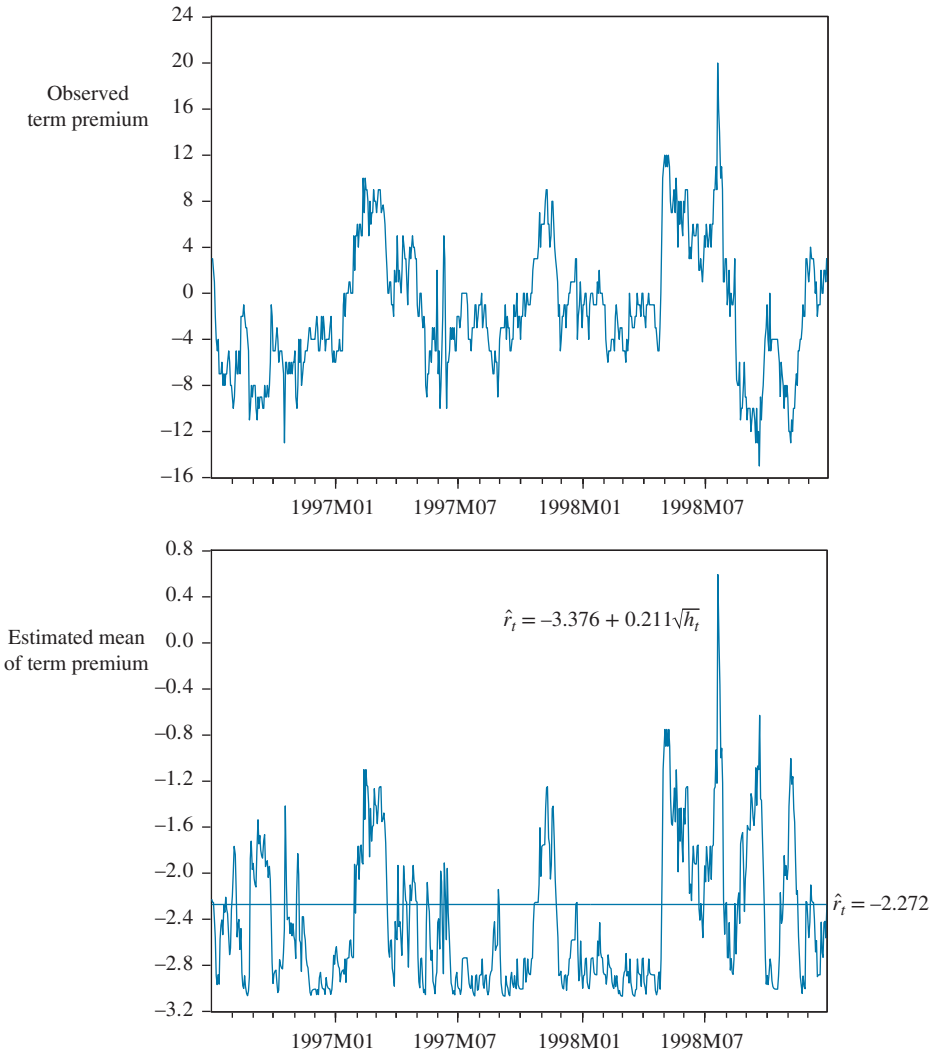


FIGURE 14.12 **Graphs for Exercise 14.11.**

a. Estimate a GARCH model and check that you obtain the following results:

$$\hat{r}_t = -2.272 \qquad \hat{h}_t = 1.729 + 0.719\hat{e}_{t-1}^2 + 0.224\hat{h}_{t-1}$$
$$(t) \qquad\qquad\qquad (6.271)\ (6.282) \qquad (3.993)$$

b. Estimate a GARCH-in-mean model and check that you obtain the following results:

$$\hat{r}_t = -3.376 + 0.211\sqrt{h_t} \qquad \hat{h}_t = 1.631 + 0.730\hat{e}_{t-1}^2 + 0.231\hat{h}_{t-1}$$
$$(t) \qquad\qquad (2.807) \qquad\qquad (5.333)\ (6.327) \qquad (4.171)$$

What is the contribution of volatility to the term premium?

c. Is the GARCH-in-mean model better than the GARCH model in a financial econometric sense? [*Hint*: Look at the statistical tests for significance.] Is the GARCH-in-mean model better than the GARCH model in a financial economic sense? [*Hint*: Look at the implications of the results, in particular the behavior of the term premium.] A plot of the expected term premium estimated for parts (a) and (b) is shown in Figure 14.12.

14.12 The data file *gold* contains 200 daily observations on the returns to shares in a company specializing in gold bullion for the period December 13, 2005, to September 19, 2006.

a. Plot the returns data. What do you notice about the volatility of returns? Identify the periods of big changes and the periods of small changes.

b. Generate the histogram of returns. Is the distribution of returns normal? Is this the unconditional or conditional distribution?

c. Perform a LM test for the presence of first-order ARCH.

d. Estimate a GARCH(1, 1) model. Are the coefficients of the correct sign and magnitude?

e. How would you use the estimated GARCH(1, 1) model to improve your forecasts of returns?

14.13 The seminal paper about ARCH by Robert Engle was concerned with the variance of UK inflation. The data file *uk* contains seasonally adjusted data on the UK consumer price index (*UKCPI*) for the sample period 1957M6 to 2006M6.

a. Compute the monthly rate of inflation (y) for the sample period 1957M7 to 2006M6 using the formula

$$y_t = 100\left[\frac{UKCPI_t - UKCPI_{t-1}}{UKCPI_{t-1}}\right]$$

b. Estimate a T-GARCH-in-mean model and check that you obtain the following results:

$$\hat{y}_t = -0.407 + 1.983\sqrt{h_t}$$
$$(t) \quad (-2.862)\ (5.243)$$
$$\hat{h}_t = 0.022 + \left(0.211 - 0.221d_{t-1}\right)e_{t-1}^2 + 0.782\hat{h}_{t-1}$$
$$(t) \quad (4.697) \quad (8.952)(-8.728) \qquad (27.677)$$

c. The negative asymmetric effect (−0.221) suggests that negative shocks (such as falls in prices) reduce volatility in inflation. Is this a sensible result for inflation?

d. What does the positive in-mean effect (1.983) tell you about inflation in the UK and volatility in prices?

14.14 The data file *warner* contains daily returns to holding shares in Time Warner Inc. The sample period is from January 3, 2008 to December 31, 2008 (260 observations), and a graph of the returns appears in Figure 14.13.

a. Estimate a GARCH(1, 1) model and an ARCH(5) model. Which model would you prefer, and why?

b. What is the expected return next period? The expected volatility next period?

c. Use your preferred model to forecast next period's return and next period's volatility.

d. Do good news and bad news have the same effect on return? On volatility?

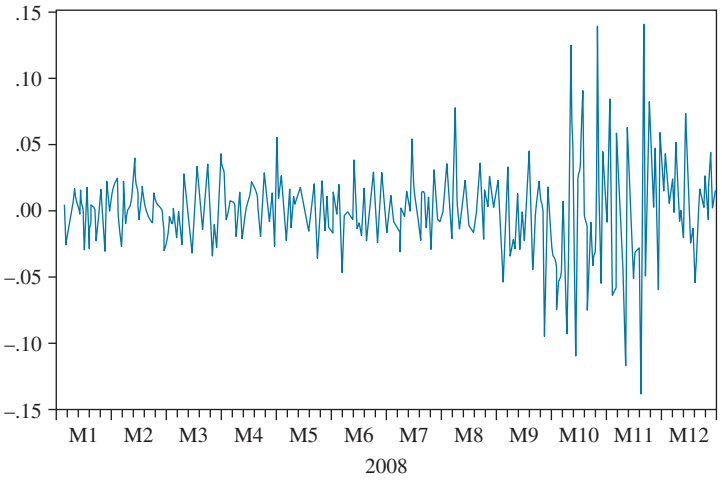


FIGURE 14.13 **Returns to shares in Time Warner.**

14.15 Consider the quarterly rates of growth contained in data file *gfc* used in Exercise 13.14. A researcher in the Euro Area (this is the group of countries in Europe where the Euro currency is the legal tender) is interested in testing the proposition that growth in the Euro region is affected by its own history, growth in the United States, and shocks to economic activity.

- **a.** Specify and estimate an econometric model for the Euro Area based only on its own history and where the expected effect of shocks on the expected quarterly rate of growth is zero.
- **b.** Specify and estimate an econometric model for the Euro Area based only on its own history and where shocks may come from distributions with zero mean, but time-varying variances.
- **c.** Specify and estimate an econometric model for the Euro Area based on its own history, the history of growth in the United States, and where the expected effect of shocks on the expected quarterly rate of growth is zero.
- **d.** Specify and estimate an econometric model for the Euro Area based on its own history and allow shocks in the Euro Area to have an effect of zero on the quarterly rate of growth.
- **e.** Specify and estimate an econometric model for the Euro Area based on its own history, the history of growth in the United States, and where shocks in the Euro Area and in the United States have an effect on the expected quarterly rate of growth.

14.16 The data file *shanghai* contains data on the daily returns to the Shanghai Stock Exchange Composite Index from July 7, 1995 to May 5, 2015.

- **a.** Plot the time series of returns and their histogram. For what observations is volatility the greatest? Describe the shape of the distribution of returns. Does the Jarque–Bera test reject the null hypothesis that returns are normally distributed?
- **b.** Estimate the GARCH model

$$y_t = \beta_0 + e_t \qquad (e_t|I_{t-1}) \sim N(0, h_t) \qquad h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$$

  Comment on the results. Plot the within-sample variance estimate $\hat{h}_t$. Have the variance estimates captured the periods of high volatility noted in part (a)?
- **c.** For the model estimated in part (b), compute the series $z_t = \hat{e}_t \Big/ \sqrt{\hat{h}_t}$. Does a histogram for the z_t suggest the assumption $z_t \sim N(0, 1)$ is valid? Does the Jarque–Bera test support this assumption?
- **d.** When the normality assumption is violated, the ordinary standard errors are not valid. However, valid robust standard errors can be used.[1] Re-estimate the model in part (b) using the Bollerslev–Wooldridge robust standard errors. Does using these standard errors change any conclusions are about the precision of estimation?

[1] See Bollerslev, T. and Wooldridge, J. (1992), "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances," *Econometric Reviews*, 11, 143–172.

e. Estimate the EGARCH model

$$y_t = \beta_0 + e_t \quad (e_t|I_{t-1}) \sim N(0, h_t) \quad \ln(h_t) = \delta + \beta_1 \ln(h_{t-1}) + \alpha \left| \frac{e_{t-1}}{\sqrt{h_{t-1}}} \right| + \gamma \left(\frac{e_{t-1}}{\sqrt{h_{t-1}}} \right)$$

Comment on the results. Plot the within-sample variance estimate $\hat{h}_t$. Have the variance estimates captured the periods of high volatility noted in part (a)?

f. For the model estimated in part (e), compute the series $z_t = \hat{e}_t \Big/ \sqrt{\hat{h}_t}$. Does a histogram for the z_t suggest the assumption $z_t \sim N(0, 1)$ is valid? Does the Jarque–Bera test support this assumption?

g. Reestimate the model in part (e) using the Bollerslev–Wooldridge standard errors. Does using these standard errors change any conclusions about the precision of estimation?

h. Find and compare estimates of $E(y_{T+1}|I_T)$ and $\text{var}(y_{T+1}|I_T)$ from the models in parts (b) and (e).

i. Using the model from part (b), and Bollerslev–Wooldridge variance and covariance estimates, find 95% interval estimates for $E(y_{T+1}|I_T)$ and $\text{var}(y_{T+1}|I_T)$.

CHAPTER 15

Panel Data Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain how a data panel differs from either a cross section or a time series of data.
2. Explain the different ways in which individual heterogeneity can be modeled using panel data, and the assumptions underlying each approach.
3. Explain how the fixed effects model allows for differences in the parameter values for each individual cross section in a data panel.
4. Compare and contrast the least squares dummy variable estimator and the fixed effects estimator.
5. Compare and contrast the fixed effects model and the random effects model. Explain what leads us to consider individual differences to be random.
6. Explain the error assumptions in the random effects model, and what characteristic leads us to consider generalized least squares estimation.
7. Describe the steps required to obtain generalized least squares estimates for the random effects estimator.
8. Explain the meaning of cluster-robust standard errors, and describe how they can be used with pooled least squares, fixed effects, and random effects estimators.
9. Explain why endogeneity is a potential problem in random effects models, and how it affects our choice of estimator.
10. Test for the existence of fixed and/or random effects, and use the Hausman test to assess whether the random effects estimator is inconsistent.
11. Explain how the Hausman–Taylor estimator can be used to obtain consistent estimates of coefficients of time-invariant variables in a random effects model.
12. Use your software to estimate fixed effects models and random effects models for panel data.

KEYWORDS

Balanced panel
Cluster-robust standard errors
Deviations about the individual mean
Difference estimator
Endogeneity
Error components model
Fixed effects estimator
Fixed effects model
Hausman test
Hausman–Taylor estimator
Heterogeneity
Instrumental variables
Least squares dummy variable model
LM test
Pooled least squares
Pooled model
Random effects estimator
Random effects model
Time-invariant variables
Time-varying variables
Unbalanced panel
Within estimator

A panel of data consists of a group of cross-sectional units (people, households, firms, states, and countries) who are observed over time. We will often refer to such units as individuals, with the term "individual" being used generically, even when the unit of interest is not a person. Let us denote the number of cross-sectional units (individuals) by N, and number of time periods in which we observe them as T. Panel data come in several different "flavors," each of which introduces new challenges and opportunities. Peter Kennedy[1] describes the different types of panel data sets as

- "Long and narrow," with "long" describing the time dimension and "narrow" implying a relatively small number of cross-sectional units
- "Short and wide," indicating that there are many individuals observed over a relatively short period of time
- "Long and wide," indicating that both N and T are relatively large

A "long and narrow" panel may consist of data on several firms over a period of time. A classic example is a data set analyzed by Grunfeld and used subsequently by many authors.[2] These data track investment in plant and equipment by $N = 11$ large firms for $T = 20$ years. This panel is narrow because it consists of only $N = 11$ firms. It is relatively "long" because $T > N$.

Many microeconomic analyses are performed on panel data sets with thousands of individuals who are followed through time. For example, the Panel Study of Income Dynamics (PSID) has followed approximately 8,000 families since 1968.[3] The U.S. Department of Labor conducts National Longitudinal Surveys (NLS) such as NLSY79, "a nationally representative sample of 12,686 young men and women who were 14–22 years old when they were first surveyed in 1979.[4] These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis." Such data sets are "wide" and "short," because N is much, much larger than T. Using panel data sets of this kind we can account for unobserved individual differences, or **heterogeneity**. Furthermore, these data panels are becoming long enough so that dynamic factors, such as spells of employment and unemployment, can be studied. These very large data sets are rich in information, and require the use of considerable computing power.

Macroeconomists who study economic growth across nations employ data that is "long" and "wide." The Penn World Table[5] provides purchasing power parity and national income accounts converted to international prices for 182 countries for some or all of the years 1950–2014, which we may roughly characterize as having both large N and large T.

Finally, it is possible to have data that combines cross-sectional and time-series data which do not constitute a panel. We may collect a sample of data on individuals from a population at several points in time, but the individuals are not the same in each time period. Such data can be used to analyze a "natural experiment," for example, when a law affecting some individuals changes, such as a change in unemployment insurance in a particular state. Using data before and after the policy change, and on groups of affected and unaffected people, the effects of the policy change can be measured. Methods for estimating effects of this type were introduced in Section 7.5.

Our interest in this chapter is how to use all available data to estimate econometric models describing the behavior of the individual cross-sectional units over time. Such data allow us to control for individual differences and study dynamic adjustment, and to measure the effects of policy changes. For each type of data, we must take care not only with error assumptions, but also

[1] *A Guide to Econometrics*, 6th ed., Chapter 18, MIT Press, 2008.

[2] See Kleiber and Zeileis, "The Grunfeld Data at 50," *German Economic Review*, 2010, 11(4), pp. 404–417 and http://statmath.wu-wien.ac.at/∼zeileis/grunfeld/.

[3] See http://psidonline.isr.umich.edu/.

[4] See www.bls.gov/nls/.

[5] See http://cid.econ.ucdavis.edu/.

with our assumptions about whether, how, and when parameters may change across individuals and/or time.

EXAMPLE 15.1 | A Microeconometric Panel

Our first example is of a data set that is short and wide. It is typical of many microeconometric analyses that use large data sets with many individuals, coming from the NLS conducted by the U.S. Department of Labor, which has a database on women who were between 14 and 24 in 1968. To illustrate, we use a subsample of $N = 716$ women who were interviewed in 1982, 1983, 1985, 1987, and 1988. The sample consists of women who were employed, and whose schooling was completed, when interviewed. The data file is named *nls_panel* and contains 3,580 lines of data. Panel data observations are usually stacked, with all the time-series observations for one individual on top of the next. The observations on a few variables for the first three women in the NLS panel are shown in Table 15.1. The first column *ID* identifies the individual and *YEAR* represents the year in which the information was collected. These identifying variables must be present so that your software will properly identify the cross-section and time-series units. Then there are observations on each of the variables. In a typical panel, there are some observations with missing values, usually denoted as "." or "NA." We have removed all the missing values in the data file *nls_panel*. In microeconomic panels, the individuals are not always interviewed the same number of times, leading to an **unbalanced panel** in which the number of time-series observations is different across individuals. The data file *nls_panel* is, however, a **balanced panel**; for each individual, we observe five time-series observations. A larger, unbalanced panel, is in the data file *nls*. Most modern software packages can handle both balanced and unbalanced panels.

TABLE 15.1 **Representative Observations from NLS Panel Data**

ID	*YEAR*	*LWAGE*	*EDUC*	*SOUTH*	*BLACK*	*UNION*	*EXPER*	*TENURE*
1	82	1.8083	12	0	1	1	7.6667	7.6667
1	83	1.8634	12	0	1	1	8.5833	8.5833
1	85	1.7894	12	0	1	1	10.1795	1.8333
1	87	1.8465	12	0	1	1	12.1795	3.7500
1	88	1.8564	12	0	1	1	13.6218	5.2500
2	82	1.2809	17	0	0	0	7.5769	2.4167
2	83	1.5159	17	0	0	0	8.3846	3.4167
2	85	1.9302	17	0	0	0	10.3846	5.4167
2	87	1.9190	17	0	0	1	12.0385	0.3333
2	88	2.2010	17	0	0	1	13.2115	1.7500
3	82	1.8148	12	0	0	0	11.4167	11.4167
3	83	1.9199	12	0	0	1	12.4167	12.4167
3	85	1.9584	12	0	0	0	14.4167	14.4167
3	87	2.0071	12	0	0	0	16.4167	16.4167
3	88	2.0899	12	0	0	0	17.8205	17.7500

15.1 The Panel Data Regression Function

A panel of data consists of a group of cross-sectional units (people, households, firms, states, or countries) who are observed over time. The sampling process we imagine is that (i) $i = 1, \ldots, N$ individuals are randomly selected from the population and (ii) each individual is observed for $t = 1, \ldots, T$ time periods. In the sampling process, we collect values y_{it} on an outcome, or dependent, variable of interest. Other characteristics concerning the individual will be used as

explanatory variables. Let $x_{1it} = 1$ be the intercept variable with $x_{2it}, \ldots, x_{Kit}$ being observations on $K - 1$ factors that vary across individual and time. Let $w_{1i}, w_{2i}, \ldots, w_{Mi}$ be observed data on M factors that do not change over time. Note that these variables **do not** have a time subscript and are said to be **time-invariant**. We cannot stress enough how important it is when using panel data to *examine the subscripts closely*, and recall that i is the indicator of the individual and t is the indicator of time.

In addition to the observed variables, there will be unobserved, omitted factors in each time period for each individual that will compose the regression's random error term. In panel data models, we can identify several types of unobserved effects. First, consider unobserved and/or unmeasurable, time-invariant individual characteristics. Let us denote these as u_{1i}, $u_{2i}, \ldots, u_{Si}$. Because we cannot observe them, we will simply refer to their combined effect as u_i, an unobserved, individual-specific random error component. Economists say that u_i represents **unobserved heterogeneity**, summarizing the unobserved factors leading to individual differences. Second, there are many, unobserved, and/or unmeasurable individual and time-varying factors $e_{1it}, e_{2it}, \ldots$ constituting the usual type of random errors in regression, and we refer to their combined effect as e_{it}. Econometricians call the random error e_{it} that varies across individual and time, an **idiosyncratic**[6] error. A third type of random error is time specific, an effect that varies over time but not individual. These factors $m_{1t}, m_{2t} \ldots$ have combined effect m_t and represent a third error component.

EXAMPLE 15.1 | Revisited

For example, in Table 15.1, the outcome variable of interest is $y_{it} = LWAGE_{it} = \ln(WAGE_{it})$. Explanatory variables include $x_{2it} = EXPER_{it}$, $x_{3it} = TENURE_{it}$, $x_{4it} = SOUTH_{it}$, and $x_{5it} = UNION_{it}$. These explanatory variables vary across both individual and time. For the indicator variables *SOUTH* and *UNION*, it means that at least some individuals moved into or out of the *SOUTH* during the 1982–1988 period, and at least some workers joined or quit a *UNION* over those years. The variables $w_{1i} = EDUC_i$ and $w_{2i} = BLACK_i$ do not change for the 716 individuals in our sample over the years 1982–1988. Two unobserved **time-invariant variables** are $u_{1i} = ABILITY_i$ and $u_{2i} = PERSEVERANCE_i$. Unobserved time-specific variables might be $m_{1t} = UNEMPLOYMENT\ RATE_t$ or $m_{2t} = INFLATION\ RATE_t$. Note that it is possible to have observable variables that change over time but not across individuals, like an indicator variable $D82_t = 1$ if the year is 1982 and $D82_t = 0$ otherwise.

A simple but representative panel data regression model is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it} \tag{15.1}$$

In (15.1), the observable outcome variable of interest is y_{it}. On the right-hand side, we have a constant term, $x_{1it} = 1$. We include one observable variable, x_{2it}, that has variation across individuals and time. The variable w_{1i} is time-invariant and varies only across individuals. The population parameters β_1, β_2, and α_1 have no subscripts and are fixed in all time periods for all individuals. We have included only one x-variable and one w-variable to keep things simple, but there can be more of each type. In parentheses, we have the two random error components, one associated with the individual (u_i) and one associated with the individual and time (e_{it}). For simplicity, we are omitting the random time-specific error component. We define the combined error

$$v_{it} = u_i + e_{it} \tag{15.2}$$

Because the regression error in (15.2) has two components, one for the individual and one for the regression, it is often called an **error components model**.

[6] Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., MIT Press, 2010, p. 285.

The complicating factor in panel data modeling is that we observe each cross-sectional unit, individual i, for more than one time-period, t. If individuals are randomly sampled, then observations on the ith individual are statistically independent of observations on the jth individual. However, using panel data, we must consider dynamic, time-related effects, and model assumptions should take them into account, just as we did in Chapter 9. The regression function of interest in a panel data model is

$$E\left[y_{it}\left|\overbrace{x_{2i1}, x_{2i2}, \ldots, x_{2iT}}^{T \text{ terms}}, w_{1i}, u_i\right.\right] = E\left(y_{it}\left|\mathbf{x}_{2i}, w_{1i}, u_i\right.\right) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i \tag{15.3}$$

where $\mathbf{x}_{2i} = \left(x_{2i1}, x_{2i2}, \ldots, x_{2iT}\right)$ represents the values x_{2it} in all time periods. Equation (15.3) says that the population average value of the outcome variable is $\beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i$, given (i) the values of x_{2it} in *all* time periods, past, present, and future; (ii) the observable individual-specific variable w_{1i}; and (iii) the unobservable individual heterogeneity term u_i. Our econometric challenge is to find a consistent and, if possible, efficient estimator for the parameters β_1, β_2, and α_1.

Equation (15.3) has several interesting features:

i. The model states that once we have controlled for x_{2it} in all time periods, and the individual-specific factors w_{1i} and u_i, only the current, contemporaneous value of x_{2it} has an effect on the expected outcome. The parameter β_2 measures the partial, or causal, effect of a change in x_{2it} on $E\left(y_{it}|\mathbf{x}_{2i}, w_{1i}, u_i\right)$, holding all else constant. Similarly, the causal effect of a change in w_{1i} on $E\left(y_{it}|\mathbf{x}_{2i}, w_{1i}, u_i\right)$ is α_1.

ii. The model conditions on the **unobservable** time-invariant error u_i. In Example 15.2, below, we examine the sales of chemical firms in China over several years using panel data. The observed explanatory variables include, for example, the amount of labor used by the firm in each year. A time-invariant variable is their location. The unobserved heterogeneity u_i might represent the ability of firm managers. The expected firm sales depend quite naturally on the unobserved managerial ability, as well as current production which depends on current labor input. However, what we are imagining is that, given managerial ability, the labor inputs of past years, or future years, have no impact on current sales.[7]

15.1.1 Further Discussion of Unobserved Heterogeneity

Every individual has unique characteristics. This is true for each of us as human beings and also for individual firms, farms, and geographic regions such as states, shires, or nations. Some individual characteristics can be observed and measured, such as an individual's height and weight, or the number of employees a firm has. Some characteristics of individuals are unmeasurable or unobservable, such as a person's ability, beauty, or fortitude. The ability of a firm's managers contributes to their revenues and profits, but just like individual ability, managerial skill is difficult or impossible to measure. Thus in a regression using cross-sectional data, these unobservable characteristics are by necessity excluded from the set of explanatory variables, and hence are included in the random error term. These unobservable individual differences are called **unobservable heterogeneity** in the economics and econometrics literature. When using panel data, it is important to separate out this component of the random error term from other components **if** we can argue that the factors causing the individual differences are unchanging over time. Such an argument is more feasible when the panel data set is wide and short, with large N and small T, as in many microeconomic panels. In a wage equation, for example, we would have to assume that unobservable factors such as ability and perseverance are constant over the period of the sample. If the panel

[7] For more discussion on this assumption, see Wooldridge (2010), p. 288.

data sample covers three or four years, we might be very comfortable with this assumption, but if the sample period covers 25 years, then we may worry about the validity of such an assumption.

Our concern with unobserved heterogeneity is exactly the same as with **omitted variables** discussed in Section 6.3.2. If omitted variables are correlated with any explanatory variables in the regression model, then the ordinary least squares (OLS) estimator suffers from **omitted variables bias**. And unfortunately, this bias does not disappear even in large samples so that the OLS estimator is inconsistent. In Chapter 10, we addressed this problem by finding a new estimator, the **instrumental variables** (IV), **two-stage least squares** (2SLS) estimator. As we will see, the beauty of having panel data is that we can control for the omitted variables bias, caused by time-invariant omitted variables, **without** having to find and use instrumental variables.

15.1.2 The Panel Data Regression Exogeneity Assumption

For the regression model (15.1), $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it})$, to have the conditional expectation in (15.3), what must be true about the random error? A new exogeneity assumption takes into account the presence of the unobserved heterogeneity term. It is

$$E(e_{it}|\mathbf{x}_{2i}, w_{i1}, u_i) = 0 \tag{15.4}$$

The meaning of this **strict exogeneity** assumption is that given the values of the explanatory variable x_{2it} in all time periods, given w_{i1} and given the unobserved heterogeneity term u_i, the best prediction of the idiosyncratic errors is zero. Another way to say this is that there is no information in these factors about the value of the idiosyncratic random error e_{it}. One subtle but extremely important point about assumption (15.4) is that it **does not** require that the unobservable heterogeneity u_i be uncorrelated with the values of the explanatory variables. We will have much more discussion about this point as we go along. Two of the implications of assumption (15.4) are that

$$\text{cov}(e_{it}, x_{2is}) = 0, \text{ and } \text{cov}(e_{it}, w_{1i}) = 0 \tag{15.5a}$$

The first part, $\text{cov}(e_{it}, x_{2is}) = 0$, is much stronger than the usual sort of exogeneity assumption. It is stronger because it is more than just contemporaneous exogeneity $\text{cov}(e_{it}, x_{2it}) = 0$; it says e_{it} is uncorrelated with *all* the values $x_{2i1}, x_{2i2}, \ldots, x_{2iT}$. In thinking about whether (15.4) is valid in a specific application, ask yourself whether (15.5a) holds. If (15.5a) is not true, and if e_{it} is correlated with any $x_{2i1}, x_{2i2}, \ldots, x_{2iT}$ or w_{1i}, then assumption (15.4) fails and the regression function of interest (15.3) is not correct.

While we are being a bit lax about it, (15.4) should properly include the intercept variable $x_{1it} = 1$, so that really $E(e_{it}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, w_{1i}, u_i) = 0$. This is important because it means that (15.5a) holds also for the intercept,

$$\text{cov}(e_{it}, x_{1is} = 1) = E(e_{it}x_{1is}) = E(e_{it}) = 0 \tag{15.5b}$$

Thus, the expected value of the idiosyncratic error is zero.

We are postponing new assumptions about error variances and covariances until Section 15.3.

15.1.3 Using OLS to Estimate the Panel Data Regression

Using our panel of data, can we consistently estimate the panel data regression function parameters in (15.3) using OLS? As we learned in Section 5.7.3 the answer is yes, if in (15.1) the **combined** error v_{it} is uncorrelated with the explanatory variable x_{2it} and with w_{1i}. That is, if

$$\text{cov}(x_{2it}, v_{it}) = E(x_{2it}v_{it}) = E(x_{2it}u_i) + E(x_{2it}e_{it}) = 0$$

and

$$\text{cov}(w_{1i}, v_{it}) = E(w_{1i}v_{it}) = E(w_{1i}u_i) + E(w_{1i}e_{it}) = 0$$

These equations say that the two random error components must be contemporaneously uncorrelated with the time-varying explanatory variables, and uncorrelated with the time-invariant explanatory variables. They in turn require

$$E(x_{2it}e_{it}) = 0, \quad E(w_{1i}e_{it}) = 0 \tag{15.6a}$$

$$E(x_{2it}u_i) = 0, \quad E(w_{1i}u_i) = 0 \tag{15.6b}$$

Equation (15.6a) says that the idiosyncratic error e_{it} is uncorrelated with the explanatory variables at time t. This is ensured by the key exogeneity assumption (15.4). On the other hand, (15.4) does not imply that (15.6b) is true, which requires the unobserved heterogeneity to be uncorrelated with the explanatory variables. The familiar example of *ABILITY* being absent from a wage equation is one case where this assumption is violated, as *ABILITY* is correlated with years of education. We should remember that if any explanatory variable is correlated with the random errors then the estimators of *all* model parameters are inconsistent. In the next section, we will introduce panel data estimation strategies that yield consistent estimators even when (15.6b) fails.

We note in passing that the model intercept variable $x_{1it} = 1$, which is exogenous, satisfies (15.6a) and (15.6b), implying that

$$E(e_{it}) = E(u_i) = E(v_{it}) = 0 \tag{15.6c}$$

Each of the random errors has mean zero. Finally, even if equations (15.6a) to (15.6c) hold, using the OLS estimator will require using a type of robust standard error which we explore in Section 15.3.

15.2 The Fixed Effects Estimator

In this section, we consider estimation procedures that employ a transformation to eliminate the individual heterogeneity from the estimation equation and thus solve the common **endogeneity** problem caused by correlation between unobservable individual characteristics and the explanatory variables. The methods achieve the same outcome using similar but different strategies. The estimators we will consider are (i) the **difference estimator**, (ii) the **within estimator**, and (iii) the **fixed effects estimator**. For each of the estimators to be consistent, the strict exogeneity assumption (15.4) must hold, but we **do not require** the unobserved heterogeneity u_i to be uncorrelated with the explanatory variables, that is, equation (15.6b) does not need to hold. The estimators successfully estimate parameters of variables that vary over time but they cannot estimate parameters of time-invariant variables. In equation (15.1), $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it}$, using these methods, we can consistently estimate β_2, but we cannot estimate β_1 or α_1.

15.2.1 The Difference Estimator: $T = 2$

It is easy to illustrate the power of having panel data with as few as $T = 2$ observations per individual, that is, when we observe each individual in two different time periods, $t = 1$ and $t = 2$. The two observations written out as in (15.1) are

$$y_{i1} = \beta_1 + \beta_2 x_{2i1} + \alpha_1 w_{1i} + u_i + e_{i1} \tag{15.7a}$$

$$y_{i2} = \beta_1 + \beta_2 x_{2i2} + \alpha_1 w_{1i} + u_i + e_{i2} \tag{15.7b}$$

Subtracting (15.7a) from (15.7b) creates a new equation

$$(y_{i2} - y_{i1}) = \beta_2 (x_{2i2} - x_{2i1}) + (e_{i2} - e_{i1}) \tag{15.8}$$

Note that (15.8) has no intercept, β_1, because it has been subtracted out. Also, $\alpha_1 w_{1i}$ subtracts out meaning that we cannot estimate the coefficient α_1 using this approach. Importantly, the unobservable individual differences u_i have dropped out due to the subtraction. Why? Because the terms β_1, $\alpha_1 w_{1i}$, and u_i are not different for time periods one and two; they are time-invariant and the subtraction removes them. We discussed variables such as $(y_{i2} - y_{i1})$ in Chapter 9. It is the change in the outcome variable's value for individual i from time period $t = 1$ to time period

$t = 2$. In the notation of Chapter 9, let "Δ" stand for "the change in," so that $\Delta y_i = (y_{i2} - y_{i1})$. Similarly, let $\Delta x_{i2} = (x_{2i2} - x_{2i1})$ and $\Delta e_i = (e_{i2} - e_{i1})$. Then, equation (15.8) becomes

$$\Delta y_i = \beta_2 \Delta x_{i2} + \Delta e_i \tag{15.9}$$

Note that a parameter of interest, β_2, is present in the transformed model (15.9). Do not be concerned about complicated data manipulations as econometric software has automatic commands to handle the differencing process.

The OLS estimator of β_2 in (15.9) is called the **first-difference estimator**, or simply the **difference estimator**. It is a consistent estimator if (i) Δe_i has zero mean and is uncorrelated with Δx_{i2}, and (ii) Δx_{i2} takes more than two values. The first condition holds if strict exogeneity, equation (15.4), holds. Recall that (15.4) implies that equations (15.5a) and (15.5b) are true. Then, Δe_i has zero mean using (15.5b). Also Δe_i is uncorrelated with Δx_{i2} because of (15.5a); the idiosyncratic error e_{it} is uncorrelated with x_{2is} in all time periods. In equation (15.8), this means that $\Delta x_{i2} = (x_{2i2} - x_{2i1})$ will be uncorrelated with $\Delta e_i = (e_{i2} - e_{i1})$.

In basic panel data analysis, the difference estimator is usually not used. We introduce it to illustrate that we can eliminate the unobserved heterogeneity through a transformation. In practice, we usually use the equivalent, but more flexible, fixed effects estimator, which we explain in Section 15.2.2.

EXAMPLE 15.2 | Using $T = 2$ Differenced Observations for a Production Function

The data file *chemical2* contains data on $N = 200$ chemical firms' sales in China for the years 2004–2006. We wish to estimate the log-log model

$$\ln(SALES_{it}) = \beta_1 + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + u_i + e_{it}$$

Using only data from 2005 and 2006, the OLS estimates with conventional, nonrobust, standard errors are

$$\widehat{\ln(SALES_{it})} = \underset{(\text{se})}{5.8745} + \underset{(0.2107)}{0.2536} \ln(CAPITAL_{it}) + \underset{(0.0354)}{0.4264} \ln(LABOR_{it}) \quad (0.0577)$$

We may be concerned that there are unobserved individual differences among the firms that are correlated with their usage of capital and labor in the production and sales process. The estimated first-difference model is

$$\widehat{\Delta \ln(SALES_{it})} = \underset{(\text{se})}{0.0384} \Delta \ln(CAPITAL_{it}) + \underset{(0.0507)}{0.3097} \Delta \ln(LABOR_{it}) \quad (0.0755)$$

There is a remarkable reduction in the estimated effect of the capital stock, which is no longer statistically significant. The estimated effect of labor is smaller but still significantly different from zero. The difference estimator is consistent when unobserved heterogeneity is correlated with the explanatory variables, but the OLS estimator is not. Given the substantial difference in the estimates we might suspect that the OLS estimates are unreliable.

EXAMPLE 15.3 | Using $T = 2$ Differenced Observations for a Wage Equation

Table 15.1 illustrates a panel data set with 5 years of data on 716 women. Consider only the final 2 years of data, 1987 and 1988, so that we have $N \times T = 716 \times 2 = 1{,}432$ observations. We wish to estimate

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_{it} + u_i + e_{it}$$

for $i = 1, \ldots, N = 716$. Note that $EDUC_i$ has no time subscript. In this sample, all the women had completed their education by the time they were first interviewed, and therefore $EDUC_i$ is time-invariant. As usual we are concerned about omitted variable bias in this model because a person's ability is unobservable. In this panel, data model ability is captured in the individual heterogeneity term u_i. Subtracting the 1987 observation from the 1988 observation, we have

$$\Delta \ln(WAGE_i) = \beta_3 \Delta EXPER_i + \Delta e_i$$

The variable *EDUC* falls out of the model because it does not take at least two values. Using the first-difference

estimator eliminates any time-invariant variables and the intercept. The change in the log of wage is attributed to the change in experience. There is no omitted variable bias because the individual heterogeneity term, which includes ability, has subtracted out. It does not matter that ability might be correlated with years of education! Using data file *nls_panel2*, the OLS estimated first difference model is

$$\underset{\text{(se)}}{\widehat{\Delta \ln(WAGE_i)}} = \underset{(0.007141)}{0.0218} \Delta EXPER_i$$

15.2.2 The Within Estimator: $T = 2$

An alternative subtraction strategy is similar in spirit to that in equation (15.8). The advantage of the **within transformation** is that it generalizes nicely to situations when we have more than $T = 2$ time observations on each individual. We begin with the models for the two time periods in (15.7a) and (15.7b), then we find the time-average of the equations, that is,

$$\frac{1}{2}\sum_{t=1}^{2}\left(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}\right)$$

On the left-hand side, we obtain $\bar{y}_{i\cdot} = \left(y_{i1} + y_{i2}\right)/2$. The "$\cdot$" is in the place of the second subscript t to remind us that it is an average over the time dimension. On the right-hand side, we obtain $\beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot}$, where the averaged variables are similarly defined: $\bar{x}_{2i\cdot} = \left(x_{2i1} + x_{2i2}\right)/2$ and $\bar{e}_{i\cdot} = \left(e_{i1} + e_{i2}\right)/2$. Note that the averaging does not affect the model parameters or the time-invariant terms β_1, w_{1i}, and u_i. The time-averaged model for $i = 1, \ldots, N$ is

$$\bar{y}_{i\cdot} = \beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot} \tag{15.10}$$

The within transformation subtracts (15.10) from the original observations to obtain

$$y_{it} - \bar{y}_{i\cdot} = \beta_2\left(x_{2it} - \bar{x}_{2i\cdot}\right) + \left(e_{it} - \bar{e}_{i\cdot}\right) \tag{15.11}$$

Instead of first-differenced variables, we have differences from the variable means. The time-invariant terms subtract out, including the unobservable heterogeneity term. Again do not be concerned about complicated data manipulations as econometric software has automatic commands to handle the process.

Let the transformed variables be denoted $\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}$, $\tilde{x}_{2it} = \left(x_{2it} - \bar{x}_{2i\cdot}\right)$, with transformed error $\tilde{e}_{it} = \left(e_{it} - \bar{e}_{i\cdot}\right)$. The within-transformed model is

$$\tilde{y}_{it} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it} \tag{15.12}$$

The OLS estimator of β_2 using (15.12) is called the **within estimator**. It is a consistent estimator if (i) $\tilde{e}_{it}$ has zero mean and is uncorrelated with $\tilde{x}_{2it}$, and (ii) if $\tilde{x}_{2it}$ takes more than two values. The first condition is satisfied if (15.4) holds. Note that the variable $\tilde{x}_{2it} = \left(x_{2it} - \bar{x}_{2i\cdot}\right)$ incorporates the values of x_{2it} in all time periods because of the average term. Similarly $\tilde{e}_{it} = \left(e_{it} - \bar{e}_{i\cdot}\right)$ depends on the values of the idiosyncratic error in all time periods because of its average. Thus, strict exogeneity, equation (15.4) is required for consistent estimation of (15.12) by OLS. Once again there is no requirement that the unobserved heterogeneity u_i be uncorrelated with the explanatory variables.

EXAMPLE 15.4 | Using the Within Transformation with $T = 2$ Observations for a Production Function

Consider using the within transformation to the $T = 2$ sales observations in Example 15.2, to estimate the effect of changes in the capital stock and labor inputs on sales. To understand the within transformation precisely, examine the transformed data on *SALES* for the first two firms in Table 15.2. For 2005 the first difference of ln(*SALES*)

is missing, which is represented by a period, "·". The time-average of the 2 year $\ln(SALES)$ is $\overline{\ln(SALES_{it})}$, and the within transformation is $\ln(SALES_{it})$. The **within** estimator uses only variation for each individual (within each individual) about the individual mean in order to estimate the parameters; it does not use variation across or between individuals in the estimation process.

There is no omitted variable bias using the within-transformed data because the time-invariant individual heterogeneity term, which includes any unmeasured characteristics of the firm, has subtracted out. Using the $N \times T = 200 \times 2 = 400$ observations the within estimates are

$$\begin{aligned} \ln(SALES_{it}) &= 0.0384 \ln(CAPITAL_{it}) \\ (\text{se}) \quad & (0.0358) \\ (\text{se}) \quad & (0.0507) \\ &+ 0.3097 \ln(LABOR_{it}) \\ & (0.0532) \quad (\text{incorrect}) \\ & (0.0755) \quad (\text{correct}) \end{aligned}$$

Notice that the within estimates are exactly the same as the first-difference estimates in Example 15.1. When $T = 2$, they will always be the same. Using OLS estimation software yields incorrect standard errors for the within estimator. The difference arises because the estimate of the error variance used by the OLS software uses the degrees of freedom $NT - 2 = 400 - 2 = 398$. The calculation ignores the loss of $N = 200$ degrees of freedom that occurs when the variables are corrected by their sample means. The correct divisor is $NT - N - 2 = 400 - 200 - 2 = 198$. Multiply the "incorrect" standard errors from the within estimates by the correction factor

$$\sqrt{(NT-2)/(NT-N-2)} = \sqrt{398/198} = 1.41778$$

The resulting "correct" standard errors are in fact identical to the standard errors from the first-difference estimator in Example 15.2. When using proper "within estimator" software this correction will automatically be done. In Section 15.2.4, we explain that most often software "within" estimator commands are called **fixed effects** estimation. The equality of the difference estimator and within estimator, and the correct standard errors, holds when $T = 2$, but not when $T > 2$.

TABLE 15.2 **Example 15.4: Transformed Sales Data**

FIRM	*YEAR*	$\ln(SALES_{it})$	$\Delta\ln(SALES_{it})$	$\overline{\ln(SALES_{it})}$	$\widetilde{\ln(SALES_{it})}$
1	2005	10.87933	·	11.08103	−0.2017047
1	2006	11.28274	0.40341	11.08103	0.2017053
2	2005	9.313799	·	9.444391	−0.1305923
2	2006	9.574984	0.261185	9.444391	0.1305927

Remark

In practice, there is no need to use the difference estimator, which was introduced as a pedagogical device to illustrate that it is possible to eliminate unobserved heterogeneity when panel data are available. Use the software option for "fixed effects" estimation.

15.2.3 The Within Estimator: $T > 2$

The advantage of the **within transformation** and use of the **within estimator** is that they generalize nicely to situations when we have more than $T = 2$ time observations on each individual. Suppose that we have T observations on each individual. So that

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

Averaging over all time observations we have

$$\frac{1}{T}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$$

On the left-hand side, we obtain $\bar{y}_{i\cdot} = (y_{i1} + y_{i2} + \cdots + y_{iT})/T$. On the right-hand side, we obtain $\beta_1 + \beta_2\bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot}$, where the averaged variables are similarly defined: $\bar{x}_{2i\cdot} = (x_{2i1} + \cdots + x_{2iT})/T$ and $\bar{e}_{i\cdot} = (e_{i1} + \cdots + e_{iT})/T$. Note that averaging does not affect the model parameters or the time-invariant terms w_{1i} and u_i. The time-averaged model, for $i = 1, \ldots, N$, is

$$\bar{y}_{i\cdot} = \beta_1 + \beta_2\bar{x}_{2i\cdot} + \alpha_1 w_{1i} + u_i + \bar{e}_{i\cdot} \tag{15.13}$$

The within transformation subtracts (15.13) from the original observations to obtain

$$y_{it} - \bar{y}_{i\cdot} = \beta_2(x_{2it} - \bar{x}_{2i\cdot}) + (e_{it} - \bar{e}_{i\cdot}) \tag{15.14}$$

Instead of first-differenced variables, we have differences from the variable means. The time-invariant variables subtract out, including the unobservable heterogeneity term.

Let the transformed variables be denoted $\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}$, $\tilde{x}_{2it} = (x_{2it} - \bar{x}_{2i\cdot})$, with transformed error $\tilde{e}_{it} = (e_{it} - \bar{e}_{i\cdot})$. The within-transformed model is

$$\tilde{y}_{it} = \beta_2\tilde{x}_{2i} + \tilde{e}_{it} \tag{15.15}$$

The OLS estimator of β_2 in (15.15) is a consistent estimator if (i) $\tilde{e}_{it}$ has zero mean and is uncorrelated with $\tilde{x}_{2it}$, and (ii) if $\tilde{x}_{2it}$ takes more than two values. These conditions hold if the strict exogeneity assumption (15.4) holds. The usual OLS standard errors for (15.15) are not quite right but are easily corrected, as we explained in Example 15.4.

EXAMPLE 15.5 | Using the Within Transformation with $T = 3$ Observations for a Production Function

Consider using the within transformation to the $T = 3$ sales observations in the data file *chemical2*, from 2004 to 2006, for the 200 firms in Example 15.2, to estimate the effect of changes in the capital stock and labor inputs on sales. The within estimates are

$$\begin{aligned}
\widehat{\ln(SALES_{it})} &= 0.0889\widehat{\ln(CAPITAL_{it})} \\
\text{(se)} \quad & (0.0271) \\
\text{(se)} \quad & (0.0332) \\
&+ 0.3522\widehat{\ln(LABOR_{it})} \\
& (0.0413) \quad \text{(incorrect)} \\
& (0.0507) \quad \text{(correct)}
\end{aligned}$$

The incorrect standard errors are produced by OLS software using $NT - 2 = 598$ degrees of freedom when it should be $NT - N - 2 = 398$. Multiplying the incorrect standard errors by the correction factor

$$\sqrt{(NT-2)/(NT-N-2)} = \sqrt{598/398} = 1.22577$$

yields correct standard errors.

15.2.4 The Least Squares Dummy Variable Model

It turns out that the within estimator is numerically equivalent to another estimator that has long been used in empirical work and that is logically appealing. To be as general as possible, we expand our equation of interest to include more variables,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it}) \tag{15.16}$$

In this regression, there is a constant term, $x_{1it} = 1$, and $(K - 1) = K_S$ variables that vary across individuals and time, and also M variables that are time invariant. There is a new symbol, K_S, that can be thought of as the number of "slope" coefficients. This will be important below when we carry out a test for the existence of individual differences.

Unobserved heterogeneity is also controlled for by including in the panel data regression (15.16) an individual-specific indicator variable for each individual. That is, let

$$D_{1i} = \begin{cases} 1 & i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad D_{2i} = \begin{cases} 1 & i = 2 \\ 0 & \text{otherwise} \end{cases}, \ldots\ldots, \quad D_{Ni} = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases}$$

Include these N indicator variables in the regression equation (15.16) to obtain

$$y_{it} = \beta_{11}D_{1i} + \beta_{12}D_{2i} + \cdots + \beta_{1N}D_{Ni} + \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$$

In this equation there is exact collinearity. The time-invariant indicator variables sum to one, $D_{1i} + D_{2i} + \cdots + D_{Ni} = 1$. Including the indicator variables requires us to drop the now redundant constant term, $x_{1it} = 1$, the time-invariant variables, w_{1i}, w_{2i}, …, w_{Mi}, *and* the unobserved heterogeneity u_i. Doing so we are left with

$$y_{it} = \beta_{11}D_{1i} + \beta_{12}D_{2i} + \cdots + \beta_{1N}D_{Ni} + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it} \tag{15.17}$$

Equation (15.17) is called the **fixed effects model**, or sometimes the **least squares dummy variable model**. The terminology **fixed effects** estimator, which is the most commonly used name in empirical work, arises because it is *as if* we are treating individual differences u_1, u_2, …, u_N, as fixed parameters, β_{11}, β_{12}, …, β_{1N}, that we can estimate. The fixed effects estimator is the OLS estimator of (15.17) using all NT observations.

Equation (15.17) is not estimated in practice unless N is small. Using the Frisch–Waugh–Lovell Theorem, Section 5.2.5 and Exercise 15.11, it can be shown that the OLS estimates of β_2, …, β_K in (15.17), and the sum of squared residuals, are identical to the within estimates of (15.16) and thus have the same consistency property under the same assumption (15.4). We remind you again that assumption (15.4) does not require that the unobserved heterogeneity term u_i be uncorrelated with $\mathbf{X}_i$ or $\mathbf{w}_i$, where $\mathbf{X}_i$ denotes all observations on the time-varying variables and $\mathbf{w}_i$ the observations on the time-invariant observations.

Remark

To summarize, the within estimator, the fixed effects estimator and the least squares dummy variable estimator are all names for the same estimators of β_2, …, β_K in (15.17). In practice, no choice is required. Use the computer software option for "fixed effects" estimation.

Because the fixed effects estimator is simply an OLS estimator, it has the usual OLS estimator variances and covariances. Including N indicator, dummy, variables means that the number of parameters is $N + K_S$, where $K_S = (K - 1)$ is the number of slope coefficients. The usual estimator of σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2}{NT - N - K_S} \tag{15.18}$$

Testing for Unobserved Heterogeneity Testing for individual differences in the fixed effects model is a test of the joint hypothesis

$$H_0 : \beta_{11} = \beta_{12}, \ \beta_{12} = \beta_{13}, \ldots, \beta_{1,N-1} = \beta_{1N}$$
$$H_1 : \text{the } \beta_{1i} \text{ are not all equal} \tag{15.19}$$

If the null hypothesis is true, then $\beta_{11} = \beta_{12} = \beta_{13} = \cdots = \beta_{1N} = \beta_1$, where β_1 denotes the common value, and there are no individual differences and no unobserved heterogeneity. The null hypothesis is $J = N - 1$ separate equalities, $\beta_{11} = \beta_{12}$, $\beta_{12} = \beta_{13}$, and so on. If the null hypothesis is true, then the "restricted model" is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$$

Under the standard OLS assumptions, the F-test statistic is

$$F = \frac{(SSE_R - SSE_U)/(N-1)}{SSE_U/(NT - N - K_S)} \tag{15.20}$$

where SSE_U is the sum of squared residuals from the fixed effects model, and SSE_R is the sum of squared errors from the OLS regression that pools all the data, $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$. If the null hypothesis is true, the test statistic has the F-distribution with $J = N - 1$ numerator degrees of freedom and $NT - N - K_S$ denominator degrees of freedom. Using the α level of significance, we reject the null hypothesis if the test statistic value is greater than, or equal to, the $1 - \alpha$ percentile of the F-distribution, $F \geq F_{(1-\alpha,\ N-1,\ NT-N-K_S)}$. The test can be made "robust" to heteroskedasticity and serial correlation, topics that we consider in Section 15.3.

EXAMPLE 15.6 | Using the Fixed Effects Estimator with $T = 3$ Observations for a Production Function

For the Chinese chemical firm data file *chemical2*, the indicator variable model in (15.21) becomes

$$\ln(SALES_{it}) = \beta_{11}D_{1i} + \cdots + \beta_{1,200}D_{200,i} + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + e_{it}$$

The fixed effects estimates of β_2 and β_3 will be identical to the within estimates in Example 15.4, and the standard errors will be the correct ones because in this indicator variable model the degrees of freedom are the correct $NT - N - (K - 1) = 600 - 200 - 2 = 398$.

The $N = 200$ estimated indicator variable coefficients, $b_{11}, b_{12}, \ldots, b_{1N}$, may or may not be of specific interest. We include the indicator variables primarily to control for unobserved heterogeneity. If, however, we are interested in predicting the sales of a specific firm then the indicator variables become crucial. Given the estimates of β_2 and β_3, $b_{11}, b_{12}, \ldots, b_{1N}$ can be recovered using the fact that the fitted regression passes through the point of the means, just as it did in the simple regression model, that is, $\bar{y}_{i\cdot} = b_{1i} + b_2\bar{x}_{2i\cdot} + b_3\bar{x}_{3i\cdot}$, $i = 1, \ldots, N$. Reporting the estimates and their standard errors is inconvenient because N may be large. Software companies cope with this in different ways. Two popular econometric software programs, EViews and Stata, report a constant term C that is the average of the estimated coefficients on the cross-section indicator variables. For the Chinese chemical firm data, $C = N^{-1}\sum_{i=1}^{N} b_{1i} = 7.5782$.

To test the null hypothesis $H_0: \beta_{11} = \beta_{12}, \beta_{12} = \beta_{13}, \ldots, \beta_{1,N-1} = \beta_{1N}$, we use the sum of squared residuals from the fixed effects estimator, $SSE_U = 34.451469$, and from the pooled OLS regression

$$\widehat{\ln(SALES_{it})} = \underset{(\text{se})}{5.8797} + \underset{(0.1711)}{0.2732}\ln(CAPITAL_{it}) + \underset{(0.0467)}{0.3815}\ln(LABOR_{it})$$

with $SSE_R = 425.636557$. The F-statistic value is

$$F = \frac{(SSE_R - SSE_U)/(N-1)}{SSE_U/(NT - N - (K-1))} = \frac{(425.636557 - 34.451469)/199}{34.451469/(600 - 200 - 2)} = 22.71$$

Using the $\alpha = 0.01$ level of significance, $F_{(0.99,\ 199,\ 398)} = 1.32$. We reject the null hypothesis and conclude that there are individual differences in the fixed effects constant terms for these $N = 200$ firms.

15.3 Panel Data Regression Error Assumptions

In Section 15.2, we considered estimation strategies that eliminate unobservable heterogeneity, u_i, so that when it is correlated with the explanatory variables we can still consistently estimate the coefficients of variables, x_{kit}, that vary across individuals and time. In this section and the next, we

propose estimation methods for the cases in which unobservable heterogeneity, u_i, is not correlated with the explanatory variables, either the **time-varying variables**, x_{kit}, or the **time-invariant variables**, w_{mi}, so that we can use OLS estimation, or a more efficient generalized least squares estimator, GLS, called the random effects (RE) estimator. Because these estimators do not eliminate unobservable heterogeneity, u_i, from the estimation equation we must make a more complete set of assumptions than we did in Section 15.2.

Panel data model estimation and inference for the model $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it})$ are complicated by the presence of two random errors. The first, u_i, accounts for time invariant unobserved heterogeneity across individuals. The second, e_{it}, is the "usual" regression error that varies across individuals and time. To be as general as possible, we return to equation (15.16), which we repeat here for your convenience,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it}) \tag{15.16}$$

As we have done in earlier chapters, let $\mathbf{x}_{it} = (1, x_{2it}, \ldots, x_{Kit})$ represent the tth observation on all time-varying variables, plus the intercept, for an individual, and let $\mathbf{X}_i$ represent all T observations on these variables for the ith individual. Let $\mathbf{w}_i = (w_{1i}, \ldots, w_{Mi})$ represent all the time-invariant variables for the ith individual. We discussed the important exogeneity assumption (15.4) that leads to the panel data regression function in (15.3). With the more complete model specification, assumption (15.4) becomes

$$E(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0 \tag{15.21}$$

Recall that the strict exogeneity assumption in (15.21) means that neither $\mathbf{X}_i$, nor $\mathbf{w}_i$, nor u_i contain any information about the possible value of the idiosyncratic random error e_{it}.

The idiosyncratic random errors e_{it} and the unobservable heterogeneity random error u_i capture quite different effects and it is plausible to treat them as statistically independent, so that there is no correlation between them. In order for the OLS estimator of (15.16) to be unbiased a strong assumption, similar to (15.21), must hold for the unobserved heterogeneity term, u_i. If the explanatory variables $\mathbf{X}_i$ and $\mathbf{w}_i$ carry no information about random error component u_i then its best prediction is zero, meaning that

$$E(u_i|\mathbf{X}_i, \mathbf{w}_i) = 0 \tag{15.22}$$

Using the law of iterated expectations, it follows that

$$E(u_i) = 0, \quad \text{cov}(u_i, x_{kit}) = E(u_i x_{kit}) = 0, \quad \text{cov}(u_i, w_{mi}) = E(u_i w_{mi}) = 0 \tag{15.23}$$

The two assumptions (15.21) and (15.22) are sufficient to ensure that the OLS estimator is unbiased and consistent.

Remark

The verb "pool" means to combine or merge things. Consequently, econometricians talk about the combined data of all individuals in all time periods as a **pooled sample**. Then the regression equation (15.16) is a **pooled model** and if we apply OLS to this pooled model it is called **pooled least squares**, or **pooled OLS**. However, pooled OLS is nothing new; it is simply the OLS estimator applied to the combined data.

Now we ask about other assumptions, namely the random error conditional variances and covariances.

Conditional Homoskedasticity The usual homoskedasticity assumption for the idiosyncratic error e_{it} is that the conditional and unconditional variances are constant,

$$\text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = \sigma_e^2 \tag{15.24a}$$

Using the variance decomposition discussed in Appendix B.1.8, and the law of iterated expectations, it also follows that

$$\text{var}(e_{it}) = E(e_{it}^2) = \sigma_e^2 \tag{15.24b}$$

Similarly, the unobserved heterogeneity random component u_i is conditionally and unconditionally homoskedastic,

$$\text{var}(u_i) = E(u_i^2) = \sigma_u^2 \tag{15.25}$$

If all individuals are drawn from one population, then homoskedasticity of u_i seems quite reasonable. However, the homoskedasticity of e_{it} is less likely to be true, for the usual reasons.

The variance of the combined error, $v_{it} = u_i + e_{it}$, is then

$$\text{var}(v_{it}|\mathbf{X}_i, \mathbf{w}_i) = \text{var}(u_i|\mathbf{X}_i, \mathbf{w}_i) + \text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i) + 2\text{cov}(u_i, e_{it}|\mathbf{X}_i, \mathbf{w}_i)$$

Combining the two homoskedasticity assumptions and the statistical independence of u_i and e_{it}, we have

$$\text{var}(v_{it}) = E(v_{it}^2) = \sigma_v^2 = \sigma_u^2 + \sigma_e^2 \tag{15.26}$$

Conditionally Correlated When unobservable heterogeneity is recognized, the usual assumption that the errors are uncorrelated does not hold. To see this, find the covariance between the combined random errors in any two time periods,

$$\begin{aligned}\text{cov}(v_{it}, v_{is}) &= E(v_{it}v_{is}) = E\left[(u_i + e_{it})(u_i + e_{is})\right] \\ &= E(u_i^2 + u_i e_{it} + u_i e_{is} + e_{it}e_{is}) \\ &= E(u_i^2) + E(u_i e_{it}) + E(u_i e_{is}) + E(e_{it}e_{is}) \\ &= \sigma_u^2\end{aligned} \tag{15.27}$$

There is a covariance between the random errors for the ith individual for observations in any two different time periods. The correlation between the errors is

$$\rho = \text{corr}(v_{it}, v_{is}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{15.28}$$

Interestingly, the covariance and correlation are constant and take the same value whether we are considering errors one period apart, or two periods apart, or more. As long as we have a random sample of individuals, we do not need to worry about any correlation *between* individuals, so that v_{it}, and v_{js} are uncorrelated for $i \neq j$.

Because of the intra individual error correlation, caused by the unobservable heterogeneity, the OLS estimator is not BLUE, and the usual standard errors are not correct. We will address how "robust" standard errors are calculated in Section 15.3.1 and how to carry out GLS in Section 15.4.

15.3.1 OLS Estimation with Cluster-Robust Standard Errors

In the panel data, multiple regression model (15.16), under the conventional homoskedasticity and serial correlation assumptions, equations (15.24a), (15.24b), (15.25), and (15.26), we have

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_e^2$$

and

$$\text{cov}(v_{it}, v_{is}) = \sigma_u^2$$

It is possible, however, that $\text{var}(e_{it})$ changes from individual to individual and perhaps also across time. In that case, $\text{var}(e_{it}) = \sigma_{it}^2$. We will introduce a new notation to handle this new possibility. Let

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_{it}^2 = \psi_{it}^2 \tag{15.29}$$

The variance ψ_{it}^2 (ψ is the Greek letter "psi") is potentially different for each individual in each time period. This might be true even if there is no unobserved heterogeneity, $\sigma_u^2 = 0$, or if the unobserved heterogeneity has a different variance for each individual. Assumption (15.29) is perfectly general and fits all possibilities.

Next, what about possible correlations among the error terms? The covariance between the random errors v_{it} and v_{is} is

$$\begin{aligned}\text{cov}(v_{it}, v_{is}) &= E(v_{it}v_{is}) = E\left[(u_i + e_{it})(u_i + e_{is})\right] \\ &= E(u_i^2) + E(e_{it}e_{is}) \\ &= \sigma_u^2 + \text{cov}(e_{it}, e_{is})\end{aligned} \tag{15.30}$$

where we have assumed u_i and e_{it} are statistically independent, or at least uncorrelated. The term $\text{cov}(e_{it}, e_{is})$ is the covariance between the usual random error, the idiosyncratic part, for the ith individual in time period t and time period s. If there is serial correlation, or autocorrelation, in this component of error then $\text{cov}(e_{it}, e_{is}) \neq 0$. The serial correlation may be of the AR(1) form we studied in Section 9.5.3, but it could be some other pattern as well. For now, we will make the most general possible assumption, that it may differ across individuals, and may differ for each pair of time periods as well, so that $\text{cov}(e_{it}, e_{is}) = \sigma_{its}$. Then (15.26) becomes

$$\text{cov}(v_{it}, v_{is}) = \sigma_u^2 + \sigma_{its} = \psi_{its} \tag{15.31}$$

Note that (15.31) is still valid even if there is no unobserved heterogeneity, so that $\sigma_u^2 = 0$.

What are the consequences of using pooled least squares in the presence of the heteroskedasticity and correlation described by (15.29) and (15.31)? The least squares estimator is still consistent, but its standard errors are incorrect, implying hypothesis tests and interval estimates based on these standard errors will be invalid. Typically, the standard errors will be too small, overstating the reliability of the least squares estimator. Fortunately, there is a way of correcting the standard errors. We had a similar situation in Chapters 8 and 9. In Chapter 8, we saw how White's heteroskedasticity-consistent standard errors could be used for assessing the reliability of least squares estimates in a regression model with heteroskedasticity of unknown form. Least squares is not efficient in these circumstances—the GLS estimator has lower variance—but using least squares avoids the need to specify the nature of the heteroskedasticity, and if the sample is large then using least squares with White standard errors provide a valid basis for interval estimation and hypothesis testing. The Newey-West standard errors introduced in Chapter 9 served a similar function in an autocorrelated-error model. They provide a valid basis for inference using least squares estimates without the need to specify the nature of the autocorrelated-error process.

In a similar way, standard errors that are valid for the pooled least squares estimator under the assumptions in (15.29) and (15.31) can be computed. These standard errors have various names, being referred to as **panel-robust standard errors** or **cluster-robust standard errors**. The T time-series observations on individuals form the clusters of data. Deriving cluster-robust standard errors requires some difficult and tedious algebra, which we briefly describe in Appendix 15A.

Two Important Notes Now for some good news and then some not so good news. First, the good news is that cluster-robust standard errors can be used in many contexts other than with panel data. Any data containing **groups** of observations can be treated as clusters if there are

within-group correlations but no across-group correlations. If we have a large sample of firms, then the firms within the same industry might define a cluster. If we have a survey of households, we may treat geographical neighborhoods as clusters. Second, the not so good news, is that while now easily obtained, using cluster-robust standard errors is not always appropriate. In order for them to be reliable, the number of individuals N must be large relative to T, so that the panel is "short and wide." For example, if there are $N = 1000$ individuals (cross sections) and we observed each for $T = 3$ time periods, then cluster-robust standard errors should work well. In situations with few individuals (few clusters) using cluster-robust standard errors may lead to inaccurate inferences. Naturally there is a great deal of discussion about what is meant by "few." In the U.S. there are $N = 50$ states. According to Cameron and Miller[8] (page 341), "Current consensus appears to be that … 50 is enough for state-year panel data." However, when carrying out tests, the number of clusters should be treated as the sample size.

EXAMPLE 15.7 | Using Pooled OLS with Cluster-Robust Standard Errors for a Production Function

In Example 15.6, we found that there is strong evidence in favor of using the fixed effects estimator rather than the pooled OLS estimator using the Chinese chemical firm data. However, for the purpose of giving a numerical illustration of pooled OLS with and without clustering, we examine the baseline model in Example 15.2 using $N = 1000$ firms using data file *chemical3*. Table 15.3 shows the OLS estimates with conventional, heteroskedasticity robust, and cluster-robust standard errors, and t-statistic values.

Note that while the heteroskedasticity-corrected standard errors are larger than the conventional standard errors, the cluster-corrected standard errors are larger yet. Of course, the t-values become smaller with the increased standard errors.

TABLE 15.3 **Example 15.7: OLS Estimates with Alternative Standard Errors**

		Conventional		Heteroskedastic		Cluster-Robust	
	Coefficient	Std. Error	t-Value	Std. Error	t-value	Std. Error	t-Value
C	5.5408	0.0828	66.94	0.0890	62.24	0.1424	38.90
ln(*CAPITAL*)	0.3202	0.0153	20.90	0.0179	17.87	0.0273	11.72
ln(*LABOR*)	0.3948	0.0225	17.56	0.0258	15.33	0.0390	10.12

15.3.2 Fixed Effects Estimation with Cluster-Robust Standard Errors

Consider now the fixed effects estimation procedure that employs the "within" transformation shown in (15.14). The within transformation removes the unobserved heterogeneity so that only the idiosyncratic error e_{it} remains. It is possible that within the cluster of observations defining each individual cross-sectional unit there remains serial correlation and/or heteroskedasticity. Cluster-robust standard errors[9] can be applied to the data in "deviation from the cluster-mean form," as in (15.14), or the least squares dummy variable model in (15.17).

[8]Cameron, A. C., and Miller, D. L., "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, 50(2), 317–373.

[9]Interestingly, the usual White heteroskedasticity robust standard errors are not valid when $T > 2$ (Cameron and Miller, 2015, p. 352). Some panel data software will automatically use cluster-robust standard errors when any kind of robust standard errors are requested.

EXAMPLE 15.8 | Using Fixed Effects and Cluster-Robust Standard Errors for a Production Function

In Example 15.7, we estimated the production function by OLS with alternative standard errors. Here using data file *chemical3*, we obtain the fixed effects estimates using $N = 1000$ firms with conventional standard errors and cluster-robust standard errors. The cluster-robust standard errors are substantially larger than the usual standard errors. When this is the case, using the cluster-robust standard errors is recommended if N is large and T is small, like they are in this case (Table 15.4).

TABLE 15.4 **Example 15.8: Fixed Effects Estimates with Alternative Standard Errors**

		Conventional		Cluster-Robust	
	Coefficient	Std. Error	*t*-Value	Std. Error	*t*-Value
C	7.9463	0.2143	37.07	0.3027	26.25
ln(*CAPITAL*)	0.1160	0.0195	5.94	0.0273	4.24
ln(*LABOR*)	0.2689	0.0307	8.77	0.0458	5.87

15.4 The Random Effects Estimator

Panel data applications fall into one of two types. The first type of application is when the unobserved heterogeneity term u_i is correlated with one or more of the explanatory variables. In this case, we use the fixed effects (within) or difference estimators because these estimators are consistent and converge in probability to the true population parameter values as the sample size increases. These estimators deal with unobserved heterogeneity by eliminating it through a transformation, eliminating the potential endogeneity problem arising from a correlation between the unobserved heterogeneity and the explanatory variables.

The second type of application is when the unobserved heterogeneity term u_i is **not** correlated with any of the explanatory variables. In this case, we can simply use pooled OLS estimation, with robust-cluster standard errors. If for our purposes the OLS estimator is sufficiently precise, then we are done. Subsequent hypothesis tests and interval estimates are valid in large samples. If the OLS estimator is not sufficiently precise, then, providing the other assumptions hold, we can use an asymptotically more efficient feasible generalized least squares (FGLS) estimator.

The panel data regression model (15.1) with unobserved heterogeneity is sometimes called the **random effects model** because individual differences are random from the point of view of the researcher. The unobservable heterogeneity terms u_i are the **random effects**. The FGLS estimator is called the **random effects estimator**. It takes into account equation (15.27), the error covariance within the observations for each individual that arises from the unobserved heterogeneity. The use of this estimator also presumes the zero conditional mean assumptions, equations (15.4), and homoskedasticity, equation (15.26).

The minimum variance, efficient, estimator for the model is a GLS estimator. As was the case when we had heteroskedasticity or autocorrelation, we can obtain the GLS estimator in the random effects model by applying OLS to a transformed model. The transformed model, using $K = 2$ and $M = 1$ in (15.16), is

$$y_{it}^* = \beta_1 x_{1it}^* + \beta_2 x_{2it}^* + \alpha_1 w_{1i}^* + v_{it}^* \tag{15.32}$$

where the transformed variables are

$$y_{it}^* = y_{it} - \alpha \bar{y}_{i\cdot}, \qquad x_{1it}^* = 1 - \alpha, \qquad x_{2it}^* = x_{2it} - \alpha \bar{x}_{2i\cdot}, \qquad w_{1i}^* = w_{1i}(1 - \alpha) \tag{15.33}$$

The transformation parameter α is between zero and one, $0 < \alpha < 1$, and is given by

$$\alpha = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}} \tag{15.34}$$

The variables $\bar{y}_{i\cdot}$ and $\bar{x}_{2i\cdot}$ are the individual time-averaged means (15.13), and w_{1i}^* is a fraction of w_{1i}. A key feature of the random effects model is that *time-invariant variables are not eliminated*. The transformed error term is $v_{it}^* = v_{it} - \alpha\bar{v}_i$. It can be shown that the transformed error v_{it}^* has constant variance σ_e^2 and is serially uncorrelated. The proof is long and tedious, so we will not inflict it on you.[10] Because the transformation parameter α depends on the unknown variances σ_e^2 and σ_u^2, these variances need to be estimated before OLS can be applied to (15.32). Some details of how the estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ are obtained can be found in Appendix 15B. The random effects, feasible GLS, estimates are obtained by applying least squares to (15.32) with σ_e^2 and σ_u^2 replaced by $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ in (15.34). From (15.33) we can see that if $\alpha = 1$ the random effects estimator is identical to the fixed effects estimator and if $\alpha = 0$ the random effects estimator is identical to the OLS estimator. When $0 < \hat{\alpha} < 1$ the random effects estimates may be closer to the OLS estimates or the fixed effects estimates depending on the magnitude of $\hat{\alpha}$.

EXAMPLE 15.9 | Random Effects Estimation of a Production Function

To illustrate the random effects estimator, we use the data file *chemical3* from $N = 1{,}000$ Chinese chemical firms using $T = 3$ time periods. The random effects estimates of the production function are

$$\begin{aligned}
\widehat{\ln(SALES_{it})} &= 6.1718 + 0.2393\ln(CAPITAL_{it}) \\
\text{(se_fgls)} \quad & (0.1142)\ (0.0147) \\
\text{(se_clus)} \quad & (0.1428)\ (0.0221) \\
&+ 0.4140\ln(LABOR_{it}) \\
& (0.0220) \\
& (0.0327)
\end{aligned}$$

These random effects estimates are obtained using the estimated "partial-demeaning coefficient"

$$\hat{\alpha} = 1 - \frac{\hat{\sigma}_e}{\sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}} = 1 - \frac{0.3722}{\sqrt{3(0.6127) + 0.1385}} = 0.7353$$

Because $\hat{\alpha} = 0.7353$ is not close to zero or one, we see that the random effects estimates are quite different from the fixed effects estimates in Example 15.8 and also quite different from the OLS estimates in Example 15.7. Note that the cluster-robust standard errors for the random effects estimates are slightly larger than the conventional FGLS standard errors, suggesting that there may be serial correlation and/or heteroskedasticity in the overall error component e_{it}.

EXAMPLE 15.10 | Random Effects Estimation of a Wage Equation

In Table 15.1, we introduced panel data using observations from a typical microeconomic data source, the National Longitudinal Surveys (NLS). In Example 15.3, we introduced a simple wage equation and noted that in the data file *nls_panel*, all the women when first surveyed had completed their education, so that the variable *EDUC*, years of education, did not vary. This resulted in it dropping out when we applied the difference estimator. All time-invariant variables are eliminated when using the difference estimator or the fixed effects estimator. In this example, we extend the model used in Example 15.3.

Because the women in our microeconomic data panel were randomly selected from a larger population, it seems sensible to treat individual differences between the 716

[10] The details can be found in Wooldridge (2010), pp. 326–328.

women as random effects. Let us specify the wage equation to have dependent variable ln(*WAGE*) and explanatory variables years of education (*EDUC*); total labor force experience (*EXPER*) and its square; tenure in current job (*TENURE*) and its square; and indicator variables *BLACK*, *SOUTH*, and *UNION*.

The fixed and random effects estimates are given in Table 15.5 along with conventional, nonrobust standard errors and *t*-values. For the random effects estimates, we use the estimated transformation parameter

$$\hat{\alpha} = 1 - \frac{\hat{\sigma}_e}{\sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}} = 1 - \frac{0.1951}{\sqrt{5 \times 0.1083 + 0.0381}}$$
$$= 0.7437$$

Using this value to transform the data as in (15.33), then applying least squares to the transformed regression model in (15.32) yields the random effects estimates. Because the random effects estimator only partially de-means the data the time-invariant variables, *EDUC* and *BLACK*, are not eliminated. We are able to estimate the effects of years of education and race on ln(*WAGE*). We estimate that the return to education is about 7.3%, and that blacks have wages about 12% lower than whites, everything else held constant. Living in the South leads to wages about 8% lower, and union membership leads to wages about 8% higher, everything else held constant.

TABLE 15.5 **Example 15.10: Fixed and Random Effects Estimates of a Wage Equation**

	Fixed Effects			Random Effects		
Variable	**Coefficient**	**Std. Error***	***t*-Value**	**Coefficient**	**Std. Error***	***t*-Value**
C	1.4500	0.0401	36.12	0.5339	0.0799	6.68
EDUC				0.0733	0.0053	13.74
EXPER	0.0411	0.0066	6.21	0.0436	0.0064	6.86
$EXPER^2$	−0.0004	0.0003	−1.50	−0.0006	0.0003	−2.14
TENURE	0.0139	0.0033	4.24	0.0142	0.0032	4.47
$TENURE^2$	−0.0009	0.0002	−4.35	−0.0008	0.0002	−3.88
BLACK				−0.1167	0.0302	−3.86
SOUTH	−0.0163	0.0361	−0.45	−0.0818	0.0224	−3.65
UNION	0.0637	0.0143	4.47	0.0802	0.0132	6.07

*Conventional standard errors.

15.4.1 Testing for Random Effects

The magnitude of the correlation ρ in (15.28) is an important feature of the random effects model. If $u_i = 0$ for every individual, then there are no individual differences and no heterogeneity to account for. In such a case, the pooled OLS linear regression model is appropriate, and there is no need for either a fixed or a random effects model. We are assuming the error component u_i has expectation zero, $E(u_i|\mathbf{X}_i, \mathbf{w}_i) = 0$. If in addition u_i has a conditional variance of *zero*, then it is said to be a degenerate random variable; it is a constant with value equal to zero. In this case, if $\sigma_u^2 = 0$, then the correlation $\rho = 0$ and there is no random individual heterogeneity present in the data. We can test for the presence of heterogeneity by testing the null hypothesis $H_0 : \sigma_u^2 = 0$ against the alternative hypothesis $H_1 : \sigma_u^2 > 0$. If the null hypothesis is rejected, then we conclude that there are random individual differences among sample members, and that the random effects model might be appropriate. On the other hand, if we fail to reject the null hypothesis, then we have no evidence to conclude that random effects are present.

The Lagrange multiplier (LM) principle for test construction is very convenient in this case, because **LM tests** require estimation of only the restricted model that assumes that the null

hypothesis is true. If the null hypothesis is true, then $u_i = 0$ and the random effects model reduces to the usual linear regression model

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + e_{it}$$

The test statistic is based on the OLS residuals

$$\hat{e}_{it} = y_{it} - b_1 - b_2 x_{2it} - a_1 w_{1i}$$

The test statistic for balanced panels is

$$\text{LM} = \sqrt{\frac{NT}{2(T-1)}} \left\{ \frac{\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{e}_{it} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2} - 1 \right\} \tag{15.35}$$

The numerator of the first term in curly brackets differs from the denominator because it contains terms like $2\hat{e}_{i1}\hat{e}_{i2} + 2\hat{e}_{i1}\hat{e}_{i3} + 2\hat{e}_{i2}\hat{e}_{i3} + \cdots$ whose sum will not be significantly different from zero if there is no correlation over time for each individual and will reflect a positive correlation if there is one. If the sum of the cross product terms is not significant, the first term in the curly brackets is not significantly different from one, and the term in the curly brackets is not significantly different from zero. If the sum of the cross product terms is significant, then the first term in the curly brackets will be significantly greater than one and LM will be positive.

If the null hypothesis $H_0 : \sigma_u^2 = 0$ is true, that is, there are no random effects, then LM $\sim N(0, 1)$ in large samples. Thus, we reject H_0 at significance level α and accept the alternative $H_1 : \sigma_u^2 > 0$ if LM $> z_{(1-\alpha)}$, where $z_{(1-\alpha)}$ is the $100(1-\alpha)$ percentile of the standard normal $N(0, 1)$ distribution.[11] This critical value is 1.645 if $\alpha = 0.05$ and 2.326 if $\alpha = 0.01$. Rejecting the null hypothesis leads us to conclude that random effects are present.

EXAMPLE 15.11 | Testing for Random Effects in a Production Function

Using the $N = 1000$ Chinese chemical firms data from *chemical3*, the value of the test statistic in (15.35) is $LM = 44.0637$. This is far greater than the $\alpha = 0.01$ critical value 2.326, so we reject the null hypothesis $H_0 : \sigma_u^2 = 0$ and conclude that $\sigma_u^2 > 0$; there is evidence of unobserved heterogeneity, or random effects, in the data.

15.4.2 A Hausman Test for Endogeneity in the Random Effects Model

The random effects model has one critical assumption that is often violated. If the random error $v_{it} = u_i + e_{it}$ is correlated with any of the right-hand side explanatory variables in a random effects model, then the least squares and GLS estimators of the parameters are biased and inconsistent.

[11] The original LM test due to Breusch and Pagan used LM^2 with the distribution under H_0 as $\chi^2_{(1)}$. Subsequent authors pointed out that the alternative hypothesis for using LM^2 is $H_1 : \sigma_u^2 \neq 0$, and that we can do better by using LM as a one-sided $N(0, 1)$ test with alternative hypothesis $H_1 : \sigma_u^2 > 0$. Some software, for example Stata, reports LM^2. The danger from using LM^2 is that LM < 0 is possible and should not be taken as evidence that $\sigma_u^2 > 0$. The adjustment for a chi-square test at significance α is to use the $100(1-2\alpha)$ percentile of the χ^2-distribution. This critical value for an $\alpha = 0.05$ test is 2.706 which is 1.645^2. It should only be used for LM > 0.

The problem of **endogenous regressors** was first considered in a general context in Chapter 10. The problem is common in random effects models because the individual-specific error component u_i may well be correlated with some of the explanatory variables. Such a correlation will cause the random effects estimator to be inconsistent. Recall that a wonderful feature of having panel data is that we can consistently estimate the model parameters using fixed effects, within, or difference estimators, without having to find instrumental variables as we did in Chapter 10. The ability to **test** whether the random effect u_i is correlated with some of the explanatory variables is important.

To check for any correlation between the error component u_i and the regressors in a random effects model, we can use a **Hausman test**. While the basic concept underlying the test is the same, the mechanics of this Hausman test are different from the Hausman test introduced in Chapter 10. In this case, the test compares the coefficient estimates from the random effects model to those from the fixed effects model. The idea underlying Hausman's test is that both the random effects and fixed effects estimators are consistent if there is no correlation between u_i and the explanatory variables x_{kit}. If both estimators are consistent, then they should converge to the true parameter values β_k in large samples. That is, in large samples, the random effects and fixed effects estimates should be similar. On the other hand, if u_i is correlated with any of the explanatory variables, then the random effects estimator is inconsistent for all the model coefficients, while the fixed effects estimator remains consistent. Thus in large samples, the fixed effects estimator converges to the true parameter values, but the random effects estimator converges to some other values that are not the values of the true parameters. In this case, we expect to see differences between the fixed and random effects estimates.

The test can be carried out coefficient by coefficient using a t-test, or jointly, using a chi-square test. Let us consider the t-test first. Denote the fixed effects estimate of β_k as $b_{FE,k}$, and let the random effects estimate be $b_{RE,k}$. Then the t-statistic for testing that there is no difference between the estimators, and thus that there is no correlation between u_i and any of the explanatory variables, is

$$t = \frac{b_{FE,k} - b_{RE,k}}{\left[\widehat{\text{var}}\left(b_{FE,k}\right) - \widehat{\text{var}}\left(b_{RE,k}\right)\right]^{1/2}} = \frac{b_{FE,k} - b_{RE,k}}{\left[\text{se}\left(b_{FE,k}\right)^2 - \text{se}\left(b_{RE,k}\right)^2\right]^{1/2}} \tag{15.36}$$

The test can be carried out for each coefficient, and if any of the t-values are statistically different from zero, then we conclude that one or more of the explanatory variables are correlated with the unobserved heterogeneity term u_i. In this t-statistic, it is important that the denominator is the estimated variance of the fixed effects estimator minus the estimated variance of the random effects estimator. The reason is that under the null hypothesis that u_i is uncorrelated with any of the explanatory variables, the random effects estimator will have a smaller variance than the fixed effects estimator, at least in large samples. Consequently, we expect to find $\widehat{\text{var}}\left(b_{FE,k}\right) - \widehat{\text{var}}\left(b_{RE,k}\right) > 0$, which is necessary for a valid test. A second interesting feature of this test statistic is that

$$\begin{aligned} \text{var}\left(b_{FE,k} - b_{RE,k}\right) &= \text{var}\left(b_{FE,k}\right) + \text{var}\left(b_{RE,k}\right) - 2\text{cov}\left(b_{FE,k}, b_{RE,k}\right) \\ &= \text{var}\left(b_{FE,k}\right) - \text{var}\left(b_{RE,k}\right) \end{aligned} \tag{15.37}$$

The unexpected result in the last line occurs because Hausman proved that, in this particular case, $\text{cov}\left(b_{FE,k}, b_{RE,k}\right) = \text{var}\left(b_{RE,k}\right)$.

More commonly, the Hausman test is automated by software packages to contrast the complete set of estimates. That is, we carry out a test of a joint hypothesis comparing all the coefficients. The Hausman contrast[12] test jointly checks how close the differences between

[12] Details of the joint test are beyond the scope of this book. A reference that contains a careful exposition of the t-test, and the chi-square test, is Wooldridge (2010), pp. 328–334.

the pairs of coefficients are to zero. When testing all the coefficients except the intercept the resulting test statistic has the $\chi^2_{(K_S)}$-distribution, where K_S is the number of coefficients of variables that vary across time and individuals, if the null hypothesis of no endogeneity is true. The form of the Hausman test in (15.36) and its χ^2-distribution equivalent are not valid for cluster-robust standard errors because under these more general assumptions it is no longer true that $\text{var}(b_{FE,k} - b_{RE,k}) = \text{var}(b_{FE,k}) - \text{var}(b_{RE,k})$.

EXAMPLE 15.12 | Testing for Endogenous Random Effects in a Production Function

Intuitively it would seem quite likely that there are unobserved characteristics of the Chinese chemical firms that might be correlated with the amount of labor and capital they use to produce their products. Let us test the differences in the coefficient β_2 of ln(*CAPITAL*) using the fixed effects estimates in Example 15.8 and the random effects estimates in Example 15.9 with conventional, nonrobust standard errors.

$$t = \frac{b_{FE,2} - b_{RE,2}}{\left[\text{se}(b_{FE,2})^2 - \text{se}(b_{RE,2})^2\right]^{1/2}} = \frac{0.1160 - 0.2393}{\left[(0.0195)^2 - (0.0147)\right]^{1/2}}$$

$$= \frac{-0.1233}{0.0129} = -9.55$$

We reject the null hypothesis that the difference in the estimators is zero, and conclude that there is endogeneity in the random effects model. Using the joint hypothesis test on the $K_S = K - 1 = 2$ coefficients yields a Hausman contrast test statistic of 98.82, which is greater than $\chi^2_{(0.95,2)} = 5.991$, leading us to conclude that there is correlation between the unobserved heterogeneity term and some of the explanatory variables. Both of these tests support the notion that in this example the random effects estimator is inconsistent, so that we should choose the fixed effects estimator for the empirical analysis.

EXAMPLE 15.13 | Testing for Endogenous Random Effects in a Wage Equation

Using the Hausman contrast test to compare the fixed and random effects estimates of the wage equation in Table 15.5 is limited to the six common coefficients. Using the individual coefficient *t*-tests you will find significant differences at the 5% level for the coefficients of $TENURE^2$, *SOUTH*, and *UNION*. The joint test for the equality of the common coefficients yields a χ^2-statistic value of 20.73 while $\chi^2_{(0.95,6)} = 12.592$. Thus both approaches lead us to conclude that there is correlation between the individual heterogeneity term and one or more of the explanatory variables and therefore the random effects estimator should not be used.

15.4.3 A Regression-Based Hausman Test

The Hausman test described in Section 15.4.2 is based on assumptions of homoskedasticity and no serial correlation. In particular, it is not robust to heteroskedasticity and/or serial correlation. A second annoying problem is that the calculated χ^2-statistic can come out to be a negative number in samples that are not large. Such a result makes no sense theoretically and is due to features of a particular sample. These problems can be avoided by using a "regression-based" Hausman test.

The test is based on an idea by Yair Mundlak, so that it is sometimes called the **Mundlak approach**. Mundlak's notion was that if the unobservable heterogeneity is correlated with the explanatory variables then perhaps the random effects are correlated with the time averages of the explanatory variables. Consider the general model in (15.16) with $K = 3$ and $M = 2$,

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + u_i + e_{it}$$

Mundlak's suggestion is that we consider

$$u_i = \gamma_1 + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i \tag{15.38}$$

where $E(c_i|\mathbf{X}_i) = 0$. Just as in the omitted variables problem, the solution is to take the relationship out of the error term and put it into the model, leaving the error with conditional expectation zero, that is, specify the panel data model

$$\begin{aligned} y_{it} &= \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + u_i + e_{it} \\ &= \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + (\gamma_1 + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i) + e_{it} \\ &= (\beta_1 + \gamma_1) + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i + e_{it} \\ &= \delta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + (c_i + e_{it}) \end{aligned} \tag{15.39}$$

Mundlak suggested testing $H_0: \gamma_2 = 0, \gamma_3 = 0$ against the alternative $H_1: \gamma_2 \neq 0$ or $\gamma_3 \neq 0$. The null hypothesis is that there is no endogeneity arising from a correlation between the unobserved heterogeneity and the explanatory variables. The asymptotically valid Wald test statistic has a $\chi^2_{(2)}$ distribution in this case. This test statistic will never be negative, and it can be made robust to heteroskedasticity and/or serial correlation using cluster-robust standard errors.

Equation (15.39) can be estimated by OLS, with cluster-robust standard errors, or by random effects, which should be more efficient. Interestingly, both OLS and random effects estimation of (15.39) yield fixed effects estimates of β_2 and β_3. Furthermore OLS and random effects estimates of γ_2 and γ_3 are identical. These outcomes are illustrated in the next two examples.

EXAMPLE 15.14 | The Mundlak Approach for a Production Function

For the production function data file *chemical3*, with $N = 1000$ firms, we create the time averages of ln(*CAPITAL*) and ln(*LABOR*) denoting them by adding a "*BAR*" over the name. The results are reported in Table 15.6. We give the estimates and standard errors to many decimal places to make the points in the previous paragraph. First, compare the OLS coefficient estimates to the random effects (RE) estimates. They are identical. Second, compare the coefficients of ln(*CAPITAL*) and ln(*LABOR*) to the fixed effects estimates in Example 15.8 and see that they are the same. Finally, note that the cluster-robust standard errors for OLS are identical to the random effects cluster-robust standard errors. The Wald test statistic value for the null hypothesis $H_0: \gamma_2 = 0, \gamma_3 = 0$ is 56.59 using cluster-robust standard errors and is 97.0 using the conventional RE standard errors. The test critical value is $\chi^2_{(0.99,2)} = 9.210$, thus using either test we reject the null hypothesis and conclude that the unobserved firm effects are correlated with the capital and/or labor inputs.

TABLE 15.6 Mundlak Regressions for a Production Function

	OLS Coefficient	Cluster Std. Error	RE Coefficient	Conventional Std. Error	RE Coefficient	Cluster Std. Error
C	5.45532814	0.14841700	5.45532814	0.13713197	5.45532814	0.14841700
ln(*CAPITAL*)	0.11603986	0.02735145	0.11603988	0.01954950	0.11603988	0.02735146
ln(*LABOR*)	0.26888033	0.04582462	0.26888041	0.03067342	0.26888041	0.04582462
$\overline{\ln(CAPITAL)}$	0.22232028	0.04125492	0.22232026	0.03338482	0.22232026	0.04125492
$\overline{\ln(LABOR)}$	0.10949491	0.06220441	0.10949483	0.05009737	0.10949483	0.06220441
Mundlak test	56.59		97.00		56.59	

EXAMPLE 15.15 | The Mundlak Approach for a Wage Equation

For the wage equation add the time averages of *EXPER* and its square, *TENURE* and its square, *SOUTH* and *UNION*. Note that we cannot use time averages of *EDUC* and *BLACK* because these variables do not change over time and are already in the model. In Table 15.7, we report the random effects estimates and both conventional and cluster-robust standard errors. The Mundlak test statistic of joint significance of the time average coefficients using the former is 20.44 and for the latter is 17.26. There are six coefficients being tested, and the test critical value is $\chi^2_{(0.99,6)} = 16.812$. Thus, we reject the null hypothesis and conclude that a woman's unobserved characteristics are correlated with some of the explanatory variables. We also for convenience provide the fixed effects (FE) estimates with cluster-robust standard errors. Note that for the time-varying variables the RE and FE coefficients are identical.

TABLE 15.7 **Mundlak Regressions for a Wage Equation**

	Random Effects			Fixed Effects	
	Coefficient	Conventional Std. Error	Cluster Std. Error	Coefficient	Cluster Std. Error
C	0.4167	0.1358	0.1101	1.4500	0.0550
EDUC	0.0708	0.0054	0.0056		
EXPER	0.0411	0.0066	0.0082	0.0411	0.0082
EXPER2	−0.0004	0.0003	0.0003	−0.0004	0.0003
TENURE	0.0139	0.0033	0.0042	0.0139	0.0042
TENURE2	−0.0009	0.0002	0.0002	−0.0009	0.0002
BLACK	−0.1216	0.0317	0.0284		
SOUTH	−0.0163	0.0361	0.0585	−0.0163	0.0585
UNION	0.0637	0.0143	0.0169	0.0637	0.0169
$\overline{EXPER}$	0.0251	0.0244	0.0223		
$\overline{EXPER2}$	−0.0012	0.0010	0.0010		
$\overline{TENURE}$	0.0026	0.0126	0.0137		
$\overline{TENURE2}$	0.0004	0.0007	0.0008		
$\overline{SOUTH}$	−0.0890	0.0464	0.0652		
$\overline{UNION}$	0.0920	0.0382	0.0415		
Mundlak test		20.44	17.26		

15.4.4 The Hausman–Taylor Estimator

The outcome from our comparison of the fixed and random effects estimates of the wage equation in Example 15.10 poses a dilemma. Correlation between the explanatory variables and the random effects means the random effects estimator will be inconsistent. We can overcome the inconsistency problem by using the fixed effects estimator, but doing so means we can no longer estimate the effects of the time-invariant variables *EDUC* and *BLACK*. The wage return for an extra year of education, and whether or not there is wage discrimination on the basis of race, might be two important questions that we would like to answer.

To solve this dilemma, we ask: How did we cope with the endogeneity problem in Chapter 10? We did so by using instrumental variable estimation. Variables known as instruments that are correlated with the endogenous variables but uncorrelated with the equation error were introduced, leading to an instrumental variables estimator which has the desirable property of consistency. The **Hausman–Taylor estimator** is an instrumental variables estimator applied

to the **random effects model**, to overcome the problem of inconsistency caused by correlation between the random effects and some of the explanatory variables. To explain how it works consider the regression model

$$y_{it} = \beta_1 + \beta_2 x_{it,exog} + \beta_3 x_{it,endog} + \beta_3 w_{i,exog} + \beta_4 w_{i,endog} + u_i + e_{it} \tag{15.40}$$

We have divided the explanatory variables into four categories:

$x_{it,exog}$: exogenous variables that vary over time and individuals

$x_{it,endog}$: endogenous variables that vary over time and individuals

$w_{i,exog}$: time-invariant exogenous variables

$w_{i,endog}$: time-invariant endogenous variables

Equation (15.40) is written as if there is one variable of each type, but in practice, there could be more than one. For the Hausman–Taylor estimator to work the number of exogenous time-varying variables $(x_{it,exog})$ must be at least as great as the number of endogenous time-invariant variables $(w_{i,endog})$. This is the necessary condition for there to be enough instrumental variables.

Following Chapter 10, we need instruments for $x_{it,endog}$ and $w_{i,endog}$. Since the fixed effects transformation $\tilde{x}_{it,endog} = x_{it,endog} - \bar{x}_{i,endog}$ eliminates correlation with u_i, we have $\tilde{x}_{it,endog}$ as a suitable instrument for $x_{it,endog}$. Also, the variables $\bar{x}_{i,exog}$ are suitable instruments for $w_{i,endog}$. The exogenous variables in (15.40) can be viewed as instruments for themselves, making the complete instrument set $x_{it,exog}$, $\tilde{x}_{it,endog}$, $w_{i,exog}$, $\bar{x}_{i,exog}$. Hausman and Taylor modify this set slightly using $\tilde{x}_{it,exog}$, $\tilde{x}_{it,endog}$, $w_{i,exog}$, $\bar{x}_{i,exog}$, which can be shown to yield the same results. Their estimator is applied to the transformed GLS model

$$y^*_{it} = \beta_1 + \beta_2 x^*_{it,exog} + \beta_3 x^*_{it,endog} + \beta_3 w^*_{i,exog} + \beta_4 w^*_{i,endog} + v^*_{it}$$

where, for example, $y^*_{it} = y_{it} - \hat{\alpha}\bar{y}_i$, and $\hat{\alpha} = 1 - \hat{\sigma}_e \Big/ \sqrt{T\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$. The estimate $\hat{\sigma}_e^2$ is obtained from fixed effects residuals; an auxiliary instrumental variables regression[13] is needed to find $\hat{\sigma}_u^2$.

EXAMPLE 15.16 | The Hausman–Taylor Estimator for a Wage Equation

For the wage equation used in Example 15.10, we will make the following assumptions

$$x_{it,exog} = \{EXPER, EXPER2, TENURE, TENURE2, UNION\}$$
$$x_{it,endog} = \{SOUTH\}$$
$$w_{i,exog} = \{BLACK\}$$
$$w_{i,endog} = \{EDUC\}$$

The variable *EDUC* is chosen as an endogenous variable on the grounds that it will be correlated with personal attributes such as ability and perseverance. It is less clear why *SOUTH* should be endogenous, but we include it as endogenous because its fixed and random effects estimates were vastly different. Perhaps those living in the South have special attributes. The remaining variables, experience, tenure, *UNION*, and *BLACK*, are assumed uncorrelated with the random effects.

Estimates for the wage equation are presented in Table 15.8. Compared to the random effects estimates, there

TABLE 15.8 **Hausman–Taylor Estimates of Wage Equation**

Variable	Coefficient	Std. Error	*t*-Value	*p*-Value
C	−0.75077	0.58624	−1.28	0.200
EDUC	0.17051	0.04446	3.83	0.000
EXPER	0.03991	0.00647	6.16	0.000
EXPER2	−0.00039	0.00027	−1.46	0.144
TENURE	0.01433	0.00316	4.53	0.000
TENURE2	−0.00085	0.00020	−4.32	0.000
BLACK	−0.03591	0.06007	−0.60	0.550
SOUTH	−0.03171	0.03485	−0.91	0.363
UNION	0.07197	0.01345	5.35	0.000

[13] Details can be found in book, Jeffrey Wooldridge (2010), pp. 358–361.

has been a dramatic increase in the estimated wage returns to education from 7.3% to 17%. The estimated effects for experience and tenure are similar. The wage reduction for *BLACK* is estimated as 3.6% rather than 11.7%, and the penalty for being in the *SOUTH* is also less, 3.1% instead of 8.2%. The instrumental-variable standard errors are mostly larger, particularly for *EDUC* and *BLACK* where the biggest changes in estimates have been observed. Which set of estimates is better will depend on how successful we have been at making the partition into exogenous and endogenous variables in (15.40) and whether the gain from having consistent estimates is sufficiently large to compensate for the increased variance of the instrumental variables estimators.

15.4.5 Summarizing Panel Data Assumptions

It will be convenient to have a summary of the assumptions under which the random effects and the fixed effects estimators are appropriate.

Random Effects Estimation Assumptions

RE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals, (ii) time-invariant variables (w_{mi}), and (iii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors, e_{it}, that vary across both time and individuals, and (ii) unobserved individual heterogeneity, u_i, that varies across individuals but not time.

RE2. (i) $E(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ and (ii) $E(u_i|\mathbf{X}_i, \mathbf{w}_i) = E(u_i) = 0$. These are the exogeneity assumptions. Condition (i) says there is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Condition (ii) says there is no information in the values of the explanatory variables that can be used to predict u_i.

RE3. (i) $\text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = \text{var}(e_{it}) = \sigma_e^2$ and (ii) $\text{var}(u_i|\mathbf{X}_i, \mathbf{w}_i) = \text{var}(u_i) = \sigma_u^2$. These are the homoskedasticity assumptions.

RE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}; (ii) the random errors e_{it} and u_i are statistically independent; and (iii) $\text{cov}(e_{it}, e_{is}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

RE5. There is no exact collinearity and all observable variables exhibit some variation.

Random Effects Estimator Notes

1. Under the assumptions RE1–RE5 the random effects (GLS) estimator is BLUE, assuming σ_e^2 and σ_u^2 are known.
2. Implementation of the random effects estimator requires the variance parameters to be estimated. The FGLS estimator is not BLUE, but it is consistent and asymptotically normal as N grows large if T is fixed, and it is asymptotically equivalent to the GLS estimator.
3. If the random errors are either heteroskedastic (RE3 fails) and/or serially correlated (RE4 (iii) fails), then the random effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect. Using cluster-robust standard errors provides a basis for valid asymptotic inference, including hypothesis tests and interval estimation.
4. Under RE1–RE5 the pooled OLS estimator is consistent and asymptotically normal.

5. Under RE1–RE5 the random effects, FGLS, estimator is more efficient asymptotically than the pooled OLS estimator with corrected cluster-robust standard errors.
6. The random effects estimator is more efficient in large samples than the fixed effects estimator for the coefficients of the variables that vary across individuals and time, x_{kit}.
7. The fixed effects estimator is, however, consistent for the coefficients of the variables that vary across individuals and time, x_{kit}, even if RE2 (ii) fails, and $E(u_i|\mathbf{X}_i, \mathbf{w}_i) \neq 0$.

Fixed Effects Estimation Assumptions

FE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + (u_i + e_{it})$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals and (ii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors e_{it} that vary across both time and individuals, (ii) unobserved individual heterogeneity u_i that varies across individuals but not time. Note that we cannot include time-invariant variables.

FE2. $E(e_{it}|\mathbf{X}_i, u_i) = 0$. This is the (strict) exogeneity assumptions. There is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Note that we do not have to make any assumption about the relationship between the unobserved heterogeneity and the explanatory variables.

FE3. $\text{var}(e_{it}|\mathbf{X}_i, u_i) = \text{var}(e_{it}) = \sigma_e^2$. The random errors e_{it} are homoskedastic.

FE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}, and (ii) $\text{cov}(e_{it}, e_{is}|\mathbf{X}_i, u_i) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

FE5. There is no exact collinearity and all observable variables exhibit some variation.

Fixed Effects Estimation Notes

1. Under FE1–FE5 the fixed effects estimator is BLUE.
2. The fixed effects estimator is consistent and asymptotically normal if N grows large and T is fixed.
3. If the random errors are either heteroskedastic (FE3 fails) and/or serially correlated (FE4 (ii) fails), then the fixed effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect. Using cluster-robust standard errors provides a basis for valid asymptotic inference, including hypothesis tests and interval estimation.

15.4.6 Summarizing and Extending Panel Data Model Estimation

The most common problem facing researchers using panel data is that unobservable characteristics of the cross-sectional unit, the "individual," are correlated with one or more of the explanatory variables. In this case, one or more of the explanatory variables are endogenous, so that OLS and the more efficient random effects estimator are inconsistent. Most of the time empirical researchers will use the fixed effects estimator because it eliminates the time-invariant unobserved heterogeneity term that causes the endogeneity problem. The fixed effects estimator is a consistent, but inefficient, estimator. Because of the major differences in the estimators, in each application using panel data, it is important to check for endogeneity using a Hausman or Mundlak test. Similarly, it is important to test for the presence of individual differences across individuals using the F-test with fixed effects estimation or the LM test for random effects.

Each of the estimators is subject to the usual problems of serial correlation and heteroskedasticity, but these problems are easily accounted for by using cluster-robust standard errors if the number of cross-sectional units N is much bigger than the time dimension T. A more perplexing problem for users of the fixed effects estimator is that time-invariant variables are eliminated from the model. In many applications, variables such as race and sex are vitally important. Using the Hausman–Taylor estimator solves the endogeneity problem by using instrumental variables estimation and does not eliminate the time-invariant variables. It can be a good choice if the IV are strong, and if there are enough time-varying exogenous variables. Another option is to use the Mundlak approach as a compromise, that is, assume that the unobserved heterogeneity depends on the time-averages of the variables varying over individual and time, as in (15.38). Once the time-averages are included in the model, if the remaining unobserved heterogeneity is not correlated with the included variables, then estimate an augmented model, like (15.39) by random effects.

Now, we briefly touch some other panel data issues.[14]

1. While we have not discussed it, panel data methods have been extended to **unbalanced panels**. These are cases when the number of time-series observations T_i differs across individuals.
2. In addition to unobserved heterogeneity associated with individuals, there can also be unobserved heterogeneity associated with time. Let m_t be a random time-specific error component. Note that the subscript is "t" only, so that it does not vary across individuals, only time. The combined error term has three terms, $v_{it} = u_i + m_t + e_{it}$. It is possible to carry out random effects estimation in this case with "two-way" error components models. A more common approach is to include a time-indicator variable in any model with relatively small T.
3. When $T = 2$, first-difference estimation is perfectly equivalent to fixed effects estimation. When $T > 2$, the first-difference random errors $\Delta v_{it} = \Delta e_{it}$ are serially correlated unless the idiosyncratic random errors e_{it} follow a random walk. This is diametrically opposite the usual fixed effects assumption that the idiosyncratic errors are serially uncorrelated. Using cluster-robust standard errors resolves the issue in both cases.
4. Dynamic panel data models that include a lagged dependent variable on the right-hand side have an endogeneity problem. To see this, let

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 y_{i,t-1} + (u_i + e_{it})$$

   Note that y_{it} depends directly on u_i, and u_i is present in every time period including time $t - 1$. Therefore, $y_{i,t-1}$ also depends directly on u_i, causing a positive correlation, making $y_{i,t-1}$ endogenous. There is large literature on this difficult problem and many innovative IV estimators have been suggested. When T is large the dynamic, time-series data characteristics, must be taken into account. Using a difference estimator in this context is very common.
5. While we have focused on endogeneity resulting from the unobserved heterogeneity term, there can be endogeneity caused by simultaneous equations, such as supply and demand equations. There are IV/2SLS methods for estimating fixed effects, RE, and first-difference models.
6. In this edition, we have chosen to omit the section on "sets of regression equations" and "seemingly unrelated regressions." These topics arise when T is large and N is small, so that each cross-sectional unit, perhaps a firm, is modeled with its own equation.[15]

[14] You are encouraged to see Badi H. Baltagi (2013) *Econometric Analysis of Panel Data, Fifth Edition*, Wiley, along with previously cited textbooks by Greene (2018) and Wooldridge (2010) for more on these topics.

[15] See Greene, pp. 328–339, or the previous edition of this book, *Principles of Econometrics*, 4th ed., 2012, Chapter 15.7.

7. Unobserved heterogeneity can affect slope coefficients, that is, it is possible that each individual's response β_{ki} to a change in x_k is different. **Random coefficient models** recognize individual-specific slopes as a possibility.[16]
8. We have mentioned the **linear probability model** for situations in which individuals face binary choices. The panel data methods we have discussed can be used with linear probability models with the usual caveats. Looking forward to Chapter 16, we introduce new estimators, probit and logit, for handing binary outcome models. These too can be adapted for panel data methods.

15.5 Exercises

15.5.1 Problems

15.1 Consider the model

$$y_{it} = \beta_{1i} + \beta_2 x_{it} + e_{it}$$

a. Show that the fixed effects estimator for β_2 can be written as

$$\hat{\beta}_{2,FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2}$$

b. Show that the random effects estimator for β_2 can be written as

$$\hat{\beta}_{2,RE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left[x_{it} - \hat{\alpha}\left(\bar{x}_i - \bar{\bar{x}}\right) - \bar{\bar{x}}\right]\left[y_{it} - \hat{\alpha}\left(\bar{y}_i - \bar{\bar{y}}\right) - \bar{\bar{y}}\right]}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left[x_{it} - \hat{\alpha}\left(\bar{x}_i - \bar{\bar{x}}\right) - \bar{\bar{x}}\right]^2}$$

where $\bar{\bar{y}}$ and $\bar{\bar{x}}$ are the overall means.

c. Write down an expression for the pooled least squares estimator of β_2. Discuss the differences between the three estimators.

15.2 Consider the panel data regression model with unobserved heterogeneity, $y_{it} = \beta_1 + \beta_2 x_{it} + v_{it} = \beta_1 + \beta_2 x_{it} + u_i + e_{it}$. Given that assumptions RE1–RE5 hold, answer each of the following questions.

a. For the purpose of estimating the regression parameters precisely by OLS, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

b. For the purpose of estimating the regression parameters precisely by GLS, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

c. For the purpose of estimating the regression parameters precisely by fixed effects, the variance of the idiosyncratic error is more important than the variance of the unobserved heterogeneity error. True or False? Explain your choice.

15.3 In the random effects model, under assumptions RE1–RE5, suppose that the variance of the idiosyncratic error is $\sigma_e^2 = \text{var}(e_{it}) = 1$.

a. If the variance of the individual heterogeneity is $\sigma_u^2 = 1$, what is the correlation ρ between $v_{it} = u_i + e_{it}$ and $v_{is} = u_i + e_{is}$?

b. If the variance of the individual heterogeneity is $\sigma_u^2 = 1$, what is the value of the GLS transformation parameter α if $T = 2$? What is the value of the GLS transformation parameter α if $T = 5$?

[16] See, for example, Greene (2018), pp. 450–459, and Wooldridge (2010), pp. 374–387.

c. In general, for any given values of σ_u^2 and σ_e^2, as the time dimension T of the panel becomes larger, the transformation parameter α becomes smaller. Is this true, false, or are you uncertain? If you are uncertain, explain.

d. If $T = 2$ and $\sigma_e^2 = \text{var}(e_{it}) = 1$, what value of σ_u^2 will give the GLS transformation parameter $\alpha = 1/4$? What value of σ_u^2 will give the GLS transformation parameter $\alpha = 1/2$? What value of σ_u^2 will give the GLS transformation parameter $\alpha = 9/10$?

e. If we think of the random errors u_i and e_{it} as noise in the regression relationship, summarize how the relative variation of these noise components, the variances of error components, affects our ability to estimate the regression parameters.

15.4 Consider the regression model $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where x_{2it} and w_{1i} are explanatory variables. The time-averaged model is given in equation (15.13), $\bar{y}_{i\cdot} = \beta_1 + \beta_2 \bar{x}_{2i\cdot} + \alpha_1 w_{1i} + \bar{v}_{i\cdot}$ where $\bar{v}_{i\cdot} = u_i + \bar{e}_{i\cdot}$. The OLS estimator of the parameters in (15.13) is called the **between estimator**, because it uses variation between, or among, individuals to estimate the regression parameters.

a. Under assumptions RE1–RE5, derive the variance of the random error $\bar{v}_{i\cdot} = u_i + \bar{e}_{i\cdot}$.

b. Under assumptions RE1–RE5, find the covariance between $\bar{v}_{i\cdot}$ and $\bar{v}_{j\cdot}$, where $i \neq j$.

c. Under assumptions RE1–RE5, the between estimator is unbiased. Is this true or false? Explain the basis of your answer.

d. If assumptions RE1–RE5 hold except for RE2, part (ii), then the between estimator is biased and inconsistent. Is this true or false? Explain the basis of your answer.

15.5 Table 15.9 contains some simulated panel data, where *id* is the individual cross-section identifier, *t* is the time period, *x* is an explanatory variable, *e* is the idiosyncratic error, *y* is the outcome value. The data generating process is $y_{it} = 10 + 5x_{it} + u_i + e_{it}$, $i = 1, 2, 3$, $t = 1, 2$. The OLS residuals are $\hat{e}$, which we have rounded to two decimal places for convenience.

TABLE 15.9 **Simulated Data for Exercises 15.5 and 15.10**

id	*t*	*x*	*e*	*y*	$\hat{e}$
1	1	−0.51	−0.69	4.43	−3.21
1	2	−0.45	−1.70	1.70	−6.31
2	1	−2.44	−0.20	−2.29	2.20
2	2	−1.26	−0.41	2.98	0.06
3	1	−0.68	0.90	11.05	4.48
3	2	1.44	1.24	22.67	2.78

a. Using the true data generating process, calculate u_i, $i = 1, 2, 3$.

b. Calculate the value of the LM statistic in equation (15.35) and carry out a test for the presence of random effects at the 5% level of significance.

c. The fixed effects estimate of the coefficient of x_{it} is $b_{FE} = 5.21$ with standard error 0.94, while the random effects estimate is $b_{RE} = 5.31$ with standard error 0.81. Test for the presence of correlation between the unobserved heterogeneity u_i and the explanatory variable x_{it}. (*Note*: The sample is actually too small for this test to be valid.)

d. If estimates of the variance components are $\hat{\sigma}_u^2 = 34.84$ and $\hat{\sigma}_e^2 = 2.59$, calculate an estimated value of the GLS transformation parameter α. Based on its magnitude, would you expect the random effects estimates to be closer to the OLS estimates or the fixed effects estimates.

e. Using the estimates in (d), compute an estimate of the correlation between $v_{i1} = u_i + e_{i1}$ and $v_{i2} = u_i + e_{i2}$. Is this correlation relatively high, or relatively low?

15.6 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. Some estimation results are in Table 15.10.

TABLE 15.10 **Estimation Results for Exercise 15.6**

	(1) OLS 1987	(2) OLS 1988	(3) FE	(4) FE Robust	(5) RE
C	0.9348	0.8993	1.5468	1.5468	1.1497
	(0.2010)	(0.2407)	(0.2522)	(0.2688)	(0.1597)
EXPER	0.1270	0.1265	0.0575	0.0575	0.0986
	(0.0295)	(0.0323)	(0.0330)	(0.0328)	(0.0220)
$EXPER^2$	−0.0033	−0.0031	−0.0012	−0.0012	−0.0023
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0007)
SOUTH	−0.2128	−0.2384	−0.3261	−0.3261	−0.2326
	(0.0338)	(0.0344)	(0.1258)	(0.2495)	(0.0317)
UNION	0.1445	0.1102	0.0822	0.0822	0.1027
	(0.0382)	(0.0387)	(0.0312)	(0.0367)	(0.0245)
N	716	716	1432	1432	1432

(standard errors in parentheses)

a. The OLS estimates of the ln(*WAGE*) model for each of the years 1987 and 1988 are reported in columns (1) and (2). How do the results compare? For these individual year estimations, what are you assuming about the regression parameter values across individuals (heterogeneity)?

b. The ln(*WAGE*) equation specified as a panel data regression model is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EXPER_{it} + \beta_3 EXPER_{it}^2 + \beta_4 SOUTH_{it} + \beta_5 UNION_{it} + (u_i + e_{it}) \quad \text{(XR15.6)}$$

Explain any differences in assumptions between this model and the models in part (a).

c. Column (3) contains the estimated fixed effects model specified in part (b). Compare these estimates with the OLS estimates. Which coefficients, apart from the intercepts, show the most difference?

d. The *F*-statistic for the null hypothesis that there are no individual differences, equation (15.20), is 11.68. What are the degrees of freedom of the *F*-distribution if the null hypothesis (15.19) is true? What is the 1% level of significance critical value for the test? What do you conclude about the null hypothesis.

e. Column (4) contains the fixed effects estimates with cluster-robust standard errors. In the context of this sample, explain the different assumptions you are making when you estimate with and without cluster-robust standard errors. Compare the standard errors with those in column (3). Which ones are substantially different? Are the robust ones larger or smaller?

f. Column (5) contains the random effects estimates. Which coefficients, apart from the intercepts, show the most difference from the fixed effects estimates? Use the Hausman test statistic (15.36) to test whether there are significant differences between the random effects estimates and the fixed effects estimates in column (3) (Why that one?). Based on the test results, is random effects estimation in this model appropriate?

15.7 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. We form first differences of the variables, such as $\Delta\ln(WAGE) = \ln(WAGE_{i,1988}) - \ln(WAGE_{i,1987})$, and specify the regression

$$\Delta\ln(WAGE) = \beta_2 \Delta EXPER + \beta_3 \Delta EXPER^2 + \beta_4 \Delta SOUTH + \beta_5 \Delta UNION + \Delta e \quad \text{(XR15.7)}$$

Table 15.11 reports OLS estimates of equation (XR15.7) as Model (1), with conventional standard errors in parentheses.

TABLE 15.11 Estimates for Exercise 15.7

Model	C	$\Delta EXPER$	$\Delta EXPER^2$	$\Delta SOUTH$	$\Delta UNION$	$SOUTH_{i,1988}$	$UNION_{i,1988}$
(1)		0.0575	−0.0012	−0.3261	0.0822		
		(0.0330)	(0.0011)	(0.1258)	(0.0312)		
(2)	−0.0774	0.1187	−0.0014	−0.3453	0.0814		
	(0.0524)	(0.0530)	(0.0011)	(0.1264)	(0.0312)		
(3)		0.0668	−0.0012	−0.3157	0.0887	−0.0220	−0.0131
		(0.0338)	(0.0011)	(0.1261)	(0.0333)	(0.0185)	(0.0231)

a. The ability of first differencing to eliminate unobservable time-invariant heterogeneity is illustrated in equation (15.8). Explain why the strict form of exogeneity, FE2, is required for the difference estimator to be consistent. You may wish to reread the start of Section 15.1.2 to help clarify the assumption.

b. Equation (XR15.6) is the panel data regression specification at the base of the difference model. Suppose we define the indicator variable $D88_t = 1$ if the year is 1988 and $D88_t = 0$ otherwise, and add it to the specification in equation (XR15.6). What would its coefficient measure?

c. Model (2) in Table 15.11 is the difference model including an intercept term. Algebraically show that the constant term added to the difference model is the coefficient of the indicator variable discussed in part (b). Is the estimated coefficient statistically significant at the 5% level? What does this imply about the intercept parameter in equation (XR15.6) in 1987 versus 1988?

d. In the difference model, the assumption of strict exogeneity can be checked. Model (3) in Table 15.11 adds the variables *SOUTH* and *UNION* for year 1988 to the difference equation. As noted in equation (15.5a), the strict exogeneity assumption fails if the random error is correlated with the explanatory variables in any time period. We can check for such a correlation by including some, or all, of the explanatory variables for year t, or $t-1$ into the difference equation. If strict exogeneity holds these additional variables should not be significant. Based on the Model (3) result is there any evidence that the strict exogeneity assumption does not hold?

e. The F-test value for the joint significance of *SOUTH* and *UNION* from part (d), in Model (3), is 0.81. Are the variables jointly significant? What are the test degrees of freedom? What is the 5% critical value?

15.8 Using the NLS panel data on $N = 716$ young women, we are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. The equation of interest is (XR15.6) in Exercise 15.6. Some estimation results are in Table 15.12. The estimates are based on 2864 observations covering the years 1982, 1983, 1985, and 1987. Standard errors are in parentheses.

TABLE 15.12 Estimates for Exercise 15.8

Model	C	$EXPER$	$EXPER^2$	$SOUTH$	$UNION$	$SOUTH_{1988}$	$UNION_{1988}$
(1)	1.3843	0.0565	−0.0011	0.0384	0.0459		
	(0.0487)	(0.0076)	(0.0003)	(0.0422)	(0.0160)		
(2)	1.3791	0.0564	−0.0011	0.0389	0.0478	0.0021	0.0160
	(0.0505)	(0.0076)	(0.0003)	(0.0451)	(0.0162)	(0.0481)	(0.0166)
robust	(0.0611)	(0.0084)	(0.0003)	(0.0636)	(0.0169)	(0.0581)	(0.0143)

a. Explain why the strict form of exogeneity, FE2, is required for the fixed effects estimator to be consistent. You may wish to reread the start of Section 15.1.2 to help clarify the assumption.

b. The fixed effects estimates of the regression coefficients and conventional standard errors are reported as Model (1). Are the coefficients significantly different from zero at the 5% level? What do the signs of the coefficients on experience and its square indicate about returns to experience?

c. In the fixed effects model, the assumption of strict exogeneity can be checked. Model (2) in Table 15.12 adds the variables *SOUTH* and *UNION* for year 1988 to the fixed effects equation and we report conventional standard and cluster-robust standard errors. As noted in equation (15.5a), the strict exogeneity assumption fails if the random error is correlated with the explanatory variables in any time period. We can check for such a correlation by including some, or all, of the explanatory variables for year $t + 1$ into the fixed effects model equation. If strict exogeneity holds these additional variables should not be significant. Based on the Model (2) result is there any evidence that the strict exogeneity assumption does not hold?

d. The joint F-test of $SOUTH_{1988}$ and $UNION_{1988}$ with conventional standard errors is 0.47. What are the degrees of freedom for the F-test? What is the 5% critical value? What do we conclude about strict exogeneity based on the joint test?

e. The joint F-test of $SOUTH_{1988}$ and $UNION_{1988}$ with robust-cluster standard errors is 0.63. When using a cluster-corrected covariance matrix the F-statistic used by some software has $M - 1$ denominator degrees of freedom, where M is the number of clusters. In this case, what is the 5% critical value? What do we conclude about strict exogeneity based on the robust joint test?

15.9 Examples 15.7 and 15.8 estimate a production function by OLS and fixed effects, respectively, with both conventional nonrobust standard errors and cluster-robust standard errors for $N = 1000$ Chinese chemical firms for 2004–2006.

a. Review the examples. What is the percent difference between the cluster-robust standard errors and the conventional standard errors?

b. Let $\hat{v}_{it}$ denote the OLS residuals from Example 15.7 and let $\hat{v}_{i,t-1}$ be the lagged residuals. Consider the regression $\hat{v}_{it} = \rho\hat{v}_{i,t-1} + r_{it}$, where r_{it} is an error term. Regressing the 2006 residuals on the 2005 residuals, we obtain $\hat{\rho} = 0.948$ with conventional OLS standard error 0.017 and White heteroskedasticity-consistent standard error 0.020. Do these results establish a time-series serial correlation in the idiosyncratic error component e_{it}? If not, what is the source of the strong correlation between $\hat{v}_{it}$ and $\hat{v}_{i,t-1}$?

c. Let $\hat{\tilde{e}}_{it}$ be the residuals from the within estimation, similar to Example 15.5, but using all 1000 firms. Let $\hat{\tilde{e}}_{i,t-1}$ be the lagged residuals. As noted in Exercise 15.10, part (e), we expect the errors in the "within" transformed model to be serially correlated with correlation $\text{corr}(\tilde{e}_{it}\tilde{e}_{is}) = -1/(T-1)$ under FE1-FE5. Here $T = 3$, thus we should find $\text{corr}(\tilde{e}_{it}\tilde{e}_{is}) = -1/2$. Consider the regression $\hat{\tilde{e}}_{it} = \rho\hat{\tilde{e}}_{i,t-1} + r_{it}$, where r_{it} is an error term. Using the 2006 data and $N = 1000$ observations, we estimate the value of ρ to be -0.233 with conventional standard error 0.046, and White heteroskedasticity robust standard error of 0.089. Test the null hypothesis $\rho = -1/2$ against the alternative $\rho \neq -1/2$ using a t-test at the 5% level, first with the conventional standard error and again with the heteroskedasticity robust standard error. Rejecting the null hypothesis implies that FE4, part (ii), does not hold, and time-series serial correlation exists in the idiosyncratic errors e_{it}. Such a finding justifies the use of cluster-robust standard errors in the fixed effects model regardless of any heteroskedasticity considerations.

d. Using the $N = 2000$ observations for 2005–2006, and the estimated regression $\hat{\tilde{e}}_{it} = \rho\hat{\tilde{e}}_{i,t-1} + r_{it}$, we estimate the value of ρ to be -0.270 with cluster-robust standard error, suggested by Wooldridge (2010, p. 311), of 0.017. Test the null hypothesis $\rho = -1/2$ against the alternative $\rho \neq -1/2$ using a t-test at the 5% level. Rejecting the null hypothesis implies that FE4, part (ii), does not hold, and time-series serial correlation exists in the idiosyncratic errors e_{it}.

15.10 This exercise uses the simulated data (y_{it}, x_{it}) in Table 15.9.

a. The fitted least squares dummy variable model, given in equation (15.17), is $\hat{y}_{it} = 5.57D_{1i} + 9.98D_{2i} + 14.88D_{3i} + 5.21x_{it}$. Compute the residuals from this estimated model for $id = 1$ and $id = 2$. What pattern do you observe in these residuals?

b. The same residual pattern occurs for $id = 3$. What is the correlation between the residuals for time periods $t = 1$ and $t = 2$?

c. The "within" model is given in equation (15.12). The transformed error is $\tilde{e}_{it} = (e_{it} - \bar{e}_{i\cdot})$. If the assumptions FE1–FE5 hold, then $\text{var}(\tilde{e}_{it}) = E\left[(e_{it} - \bar{e}_{i\cdot})^2\right]$, where $\bar{e}_{i\cdot} = (e_{i1} + e_{i2})/2$ because $T = 2$. Show that $\text{var}(\tilde{e}_{it}) = \sigma_e^2/2$.

d. Using the same approach, as in part (c), show that $\text{cov}(\tilde{e}_{i1}, \tilde{e}_{i2}) = E\left[(e_{i1} - \bar{e}_{i\cdot})(e_{i2} - \bar{e}_{i\cdot})\right] = -\sigma_e^2/2$.

e. Using the results in parts (c) and (d), it follows that $\text{corr}(\tilde{e}_{i1}, \tilde{e}_{i2}) = -1$. Relate this result to your answer in (b). In fact for $T > 1$, and assuming FE1-FE5 hold, $\text{corr}(\tilde{e}_{it}, \tilde{e}_{is}) = -1/(T-1)$ if $t \neq s$. We anticipate the within-transformed errors to be serially correlated.

15.11 Several software companies report fixed effects estimates with an estimated intercept. As explained in Example 15.6, the value they report is the average of the coefficients of the indicator variables in the least squares dummy variable model, given in equation (15.17). Using the data in Table 15.9, the fitted dummy variable model is $\hat{y}_{it} = 5.57D_{1i} + 9.98D_{2i} + 14.88D_{3i} + 5.21x_{it}$.

a. Compute the average of the dummy variable coefficients, calling it C.

b. The fitted fixed effects model, using the device from part (a), is $\hat{y}_{it} = C + 5.21x_{it}$. Calculate $\bar{y}_{i\cdot} - b_2\bar{x}_{2i\cdot}$ for $id = 1$ and $id = 2$. For your convenience, to two decimals, $\bar{y}_{1\cdot} = 3.07$, $\bar{y}_{2\cdot} = 0.34$ and $\bar{x}_{1\cdot} = -0.48$, $\bar{x}_{2\cdot} = -1.85$. Round the calculated values to two decimals and compare them to the dummy variable coefficients.

c. Given the fitted model $\hat{y}_{it} = C + 5.21x_{it}$, compute the residuals for $id = 1$ and $id = 2$.

d. What is the fitted within-model equation (15.17)?

e. Calculate the within-model residuals for $id = 1$ and $id = 2$.

f. Explain the relationship between the within model residuals in part (e) and the residuals calculated in part (c), apart from any error caused by the two decimal rounding.

15.12 Do larger universities have lower cost per student or a higher cost per student? A university is many things and here we only focus on the effect of undergraduate full-time student enrollment (*FTESTU*) on average total cost per student (*ACA*). Consider the regression model $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + e_{it}$ where the subscripts i denote the university and t refers to the time period, and e_{it} is the usual random error term.

a. Using the 2010–2011 data on 141 public universities, we estimate the model above. The estimate of β_2 is $b_2 = 0.28$. The 95% interval estimate is [0.169, 0.392]. What is the estimated effect of increasing enrollment on average cost per student? Is there a statistically significant relationship?

b. There are many other factors affecting average cost per student besides enrollment. Some of them can be characterized as the university "identity" or "image." Let us denote these largely unobservable individual characteristics attributes as u_i. If we add this feature to the model, it becomes $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + (u_i + e_{it}) = \beta_1 + \beta_2 FTESTU_{it} + v_{it}$. As long as v_{it} is statistically independent of full-time student enrollment, then the least squares estimator is BLUE. Is that true or false? Explain your answer.

c. The combined error is $v_{it} = u_i + e_{it}$. Let $\hat{v}_{it}$ be the least squares residual from the regression in (a). We then estimate a simple regression with dependent variable $\hat{v}_{i,2011}$ and explanatory variable $\hat{v}_{i,2010}$. The estimated coefficient is 0.93 and very significant. Is this evidence in support of the presence of unobservable individual attributes u_i, or against them? Explain your logic.

d. With our 2 years of data, we can take "first differences" of the model in (b). Subtracting the model in 2010 from the model in 2011, we have $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta v_i$, where

$$\Delta ACA_i = ACA_{i,2011} - ACA_{i,2010},$$

$$\Delta FTESTU_i = FTESTU_{i,2011} - FTESTU_{i,2010}$$

$$\text{and } \Delta v_i = v_{i,2011} - v_{i,2010}$$

Using the first-difference model, and given the results in (c), will there be serial correlation in the error Δv_i? Explain your reasoning.

e. Using OLS, we estimate the model in (d) and the resulting estimate of β_2 is $b_{FD} = -0.574$ with standard error $\text{se}(b_{FD}) = 0.107$. What now is the estimated effect of increasing enrollment on average cost per student? Explain why the result of this regression is so different from the pooled regression result in (a). Which set of estimates do you believe are more plausible? Why?

15.13 Consider the panel data regression in equation (15.1) for N cross-sectional units with $T = 3$ time-series observations. Assume that FE1–FE5 hold.

a. Apply the first-difference transformation to model (15.1). What is the resulting specification? Is there unobserved heterogeneity in this model? Explain.

b. Let $\Delta e_{it} = (e_{it} - e_{i,t-1})$. Find the variance of Δe_{it} for $t = 2$ and $t = 3$.

c. Assuming that the idiosyncratic error e_{it} is serially uncorrelated, show that the correlation between Δe_{i3} and Δe_{i2} is $-1/2$.

d. What must the serial correlation for e_{it} be in order for Δe_{i3} and Δe_{i2} to be uncorrelated?

15.14 Using the NLS panel data on $N = 716$ young women for years 1982, 1983, 1985, 1987, and 1988, we are interested in the relationship between ln(*WAGE*) and education, experience, its square, usual hours worked per week, and an indicator variable for black women. The equation is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_{it} + \beta_4 EXPER_{it}^2 + \beta_5 HOURS_{it} + \beta_6 BLACK_i + u_i + e_{it}$$

Table 15.13 contains OLS, random effects, and Hausman–Taylor model estimates for this model and includes conventional and cluster-robust standard errors for each. The Hausman–Taylor estimator treats *EDUC* and *HOURS* as endogenous and correlated with the unobserved heterogeneity.

TABLE 15.13 **Estimates for Exercise 15.14**

	C	*EDUC*	*EXPER*	$EXPER^2$	*HOURS*	*BLACK*
OLS	0.4509	0.0748	0.0631	−0.0012	−0.0008	−0.1347
(se)	(0.0617)	(0.0028)	(0.0080)	(0.0003)	(0.0008)	(0.0149)
(robust)	(0.1030)	(0.0055)	(0.0100)	(0.0004)	(0.0019)	(0.0290)
RE	0.6294	0.0769	0.0591	−0.0011	−0.0054	−0.1271
(se)	(0.0833)	(0.0055)	(0.0056)	(0.0002)	(0.0007)	(0.0298)
(robust)	(0.0999)	(0.0054)	(0.0069)	(0.0003)	(0.0017)	(0.0294)
HT	0.2153	0.1109	0.0583	−0.0011	−0.0063	−0.0910
(se)	(0.5536)	(0.0422)	(0.0057)	(0.0002)	(0.0007)	(0.0529)
(robust)	(0.4897)	(0.0381)	(0.0075)	(0.0003)	(0.0018)	(0.0494)

a. What is the interpretation of β_2? How much difference is there among the OLS, random effects, and Hausman–Taylor estimates of β_2? Construct a 95% interval estimate for β_2 using each estimator and cluster-robust standard errors. What differences do you observe?

b. For the Hausman–Taylor estimator, how many instrumental variables are required? How many instruments do we have? What are they?

c. For this model, why might we prefer the Hausman–Taylor estimator to the fixed effects estimator?

d. The fixed effects estimates of the coefficients of *EXPER*, $EXPER^2$, and *HOURS* and their conventional standard errors are 0.0584 (0.00574), −0.0011 (0.00023), and −0.0063 (0.00074), respectively. Comparing these estimates to the random effects estimates, with conventional standard errors, are we justified in worrying about endogeneity in this model?

e. By using cluster-robust standard errors for the random effects estimator, which of the assumptions RE1–RE5 are we relaxing?

f. Using the Hausman–Taylor model, $\hat{\sigma}_u = 0.35747$ and $\hat{\sigma}_e = 0.19384$. Given these estimates, which source of error variation is more important in this model? The variation in unobserved heterogeneity or the variation in the idiosyncratic error? What is the proportion of the combined variation that is accounted for by the unobserved heterogeneity?

15.15 Using 352 observations on 44 rice farmers in the Tarlac region of the Phillipines for 8 years from 1990 to 1997, we estimated the relationship between tonnes of freshly threshed rice produced (*PROD*), hectares planted (*AREA*), person-days of hired and family labor (*LABOR*), and kilograms of fertilizer (*FERT*). The log–log specification of the model, including the unobserved heterogeneity term, is

$$\ln(PROD_{it}) = \beta_1 + \beta_2 \ln(AREA_{it}) + \beta_3 \ln(LABOR_{it}) + \beta_4 \ln(FERT_{it}) + u_i + e_{it}$$

Table 15.14 contains various estimates of the model. Model (1) contains OLS estimates. Model (2) contains OLS estimates of the model including year dummy variables, which are not shown, such

as $D91 = 1$ for year 1991, $D91 = 0$ otherwise. Model (3) contains fixed effects estimates. Model (4) contains fixed effects estimates of the model including year dummy variables. In each case, conventional standard errors are reported, (se), and for Model (4), we also report cluster-robust standard errors (robust). For each model, we report the sum of squared residuals and the number of model parameters, apart from the intercept. The p-values are reported for the t-statistics computed using the conventional standard errors.

TABLE 15.14 Estimates for Exercise 15.15

Model		C	ln(*AREA*)	ln(*LABOR*)	ln(*FERT*)	*SSE*	*K*−*1*
(1)	OLS	−1.5468***	0.3617***	0.4328***	0.2095***	40.5654	3
	(se)	(0.2557)	(0.0640)	(0.0669)	(0.0383)		
(2)	OLS	−1.5549***	0.3759***	0.4221***	0.2075***	36.2031	10
	(se)	(0.2524)	(0.0618)	(0.0663)	(0.0380)		
(3)	FE	−0.3352	0.5841***	0.2586***	0.0952*	27.6623	46
	(se)	(0.3263)	(0.0802)	(0.0703)	(0.0432)		
(4)	FE	−0.3122	0.6243***	0.2412***	0.0890*	23.0824	53
	(se)	(0.3107)	(0.0755)	(0.0682)	(0.0415)		
	(robust)	(0.5748)	(0.0971)	(0.0968)	(0.0881)		

*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

a. Comment on the sensitivity of the estimates of the input elasticities to the various models.
b. Which of the estimated models do you prefer? Perform a series of hypothesis tests to help you make your decision.
c. For Model (4), find 95% interval estimates for the input elasticities using (i) conventional standard errors and (ii) cluster-robust standard errors. Comment on any differences.
d. Calculate the p-value for the coefficient of ln(*FERT*) using the robust standard error.

15.5.2 Computer Exercises

15.16 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*), (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Using the data on *INCOME* for the first household, calculate the time average, within and differenced observations for *INCOME*. What is the sum of the within-transformed observations on *INCOME* for the first household?
b. Consider the panel data regression model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ where $i = 1, 2, \ldots, 40$ refers to household and $t = 1, 2, 3$ refers to year. Obtain the OLS estimates of this model.
c. What are the fixed effects estimates of the parameters? What is the sum of squared residuals? Using the sum of squared residuals from the fixed effects estimates and the OLS estimation in (b), test for the presence of individual differences using an F-test. Show how the test statistic is computed. Using the 5% level of significance, what do we conclude?
d. Using OLS, regress *LIQUOR* on a constant term and 39 individual-specific indicator variables. Save the OLS residuals and call them *LIQUORW*. Regress *INCOME* on a constant term and 39 individual-specific indicator variables. Save the residuals and call them *INCOMEW*. Using OLS regress *LIQOURW* on *INCOMEW* without a constant term. What is the estimated coefficient? What is the sum of squared errors? How does this exercise illustrate the Frisch–Waugh–Lovell theorem discussed in Section 5.2.5?
e. Following Example 15.5, show how to correct the standard errors from the regression of *LIQOURW* on *INCOMEW* to make them match the fixed effects standard errors.

15.17 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*) (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Create the first-differenced observations on *LIQUOR* and *INCOME*. Call these new variables *LIQUORD* and *INCOMED*. Using OLS regress *LIQUORD* on *INCOMED* without a constant term. Construct a 95% interval estimate of the coefficient.

b. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ using random effects. Construct a 95% interval estimate of the coefficient on *INCOME*. How does it compare to the interval in part (a)?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

d. For each individual, compute the time averages for the variable *INCOME*. Call this variable *INCOMEM*. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + \gamma INCOMEM_i + c_i + e_{it}$ using the random effects estimator. Test the significance of the coefficient γ at the 5% level. Based on this test, what can we conclude about the correlation between the random effect u_i and *INCOME*? Is it OK to use the random effects estimator for the model in (b)?

15.18 The data file *mexican* contains data collected in 2001 from the transactions of 754 female Mexican sex workers. There is information on four transactions per worker.[17] The labels *ID* and *TRANS* are used to describe a particular woman and a particular transaction. There are three categories of variables.

1. Sex worker characteristics: (i) *AGE*, (ii) an indicator variable *ATTRACTIVE* equal to 1 if the worker is attractive, and (iii) an indicator variable *SCHOOL* if she has completed secondary school or higher.
2. Client characteristics: (i) an indicator variable *REGULAR* equal to 1 if the client is a regular, (ii) an indicator variable *RICH* equal to 1 if the client is rich, and (iii) an indicator variable *ALCOHOL* if the client has consumed alcohol before the transaction.
3. Transaction characteristics: (i) the log of the price of the transaction *LNPRICE*, (ii) an indicator variable *NOCONDOM* equal to 1 if a condom was not used, and (iii) two indicator variables for location, *BAR* equal to 1 if the transaction originated in bar and *STREET* equal to 1 if the transaction originated in the street.

a. Using OLS, estimate a relationship with *LNPRICE* as the dependent variable, and as explanatory variables the sex worker characteristics, client characteristics, and transaction characteristics. Discuss the signs and significance of the estimated coefficients.

b. Gertler, Shah, and Bertozzi argue that the coefficient of *NOCONDOM* is a risk premium. Some sex workers are willing to take the risk of having unprotected sex because of the extra price some clients are willing to pay to avoid using a condom. What is your 95% interval estimate of the risk premium based on these OLS estimates?

c. What are some factors that might be included in an unobserved heterogeneity error component in this model? A crucial assumption for the consistency of the OLS estimator is that the unobserved heterogeneity term is uncorrelated with the explanatory variables. Without carrying out a formal test, what are your thoughts about this exogeneity assumption for the model in (a)?

d. Estimate the model in part (a) using the fixed effects estimator, omitting sex worker characteristics. (i) Why did we omit the sex worker characteristics? and (ii) Which coefficient estimates are significantly different from zero at a 5% level of significance?

e. Using the fixed effects estimation in (d), carry out an F-test for the presence of individual sex worker differences. Use the 1% level of significance.

f. Using the fixed effects estimates, how is the price affected when clients are rich, are regular, and have consumed alcohol? How does the location of the transaction influence the price?

g. What is your 95% interval estimate of the risk premium based on these fixed effects estimates? Compare this interval estimate to the one in part (b).

15.19 This exercise uses the data and model in Exercise 15.18.

a. Estimate the model assuming random effects and with the characteristics of the sex workers included in the model. Carry out a test of the joint significance of the sex worker characteristics at the 5% level. Are these coefficients jointly significant? Are they individually significant?

[17] These data are a subset of those used by Paul Gertler, Manisha Shah and Stefano Bertozzi in their study "Risky Business: The Market for Unprotected Sex", *Journal of Political Economy*, 2005, 113, 518–550.

b. What is your 95% interval estimate of the risk premium, the coefficient on *NOCONDOM*, based on these random effects estimates?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

d. Based on the random effects estimates, how much extra does a client have to pay to have unprotected sex with an attractive secondary-educated sex worker?

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *NOCONDOM*, *RICH*, *REGULAR*, *ALCOHOL*, *BAR*, and *STREET*. If there are significant differences between any of the coefficients, should we rely on the fixed effects estimates or the random effects estimates? Explain your choice.

f. Reconsider the random effects model from part (a), but assume *NOCONDOM* is correlated with the random effects. Reestimate the model using the Hausman–Taylor estimator with *NOCONDOM* treated as endogenous. Compare the results with those obtained in part (b). How much extra does a client have to pay to have unprotected sex with an attractive secondary-educated sex worker? What is your 95% interval estimate of the risk premium, the coefficient on *NOCONDOM*, based on the Hausman–Taylor estimates?

15.20 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does the student's sex or race make a difference?

b. Reestimate the model in part (a) with school fixed effects. Compare the results with those in part (a). Have any of your conclusions changed? [*Hint*: specify *SCHID* as the cross-section identifier and *ID* as the "time" identifier.]

c. Test for the significance of the school fixed effects. Under what conditions would we expect the inclusion of significant fixed effects to have little influence on the coefficient estimates of the remaining variables?

d. Reestimate the model in part (a) with school random effects. Compare the results with those from parts (a) and (b). Are there any variables in the equation that might be correlated with the school effects? Use the LM test for the presence of random effects.

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *SMALL*, *AIDE*, *TCHEXPER*, *WHITE_ASIAN*, and *FREELUNCH*. What are the implications of the test outcomes? What happens if we apply the test to the fixed and random effects estimates of the coefficient on *BOY*?

f. Create school-averages of the variables and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

15.21 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. It replicates Exercise 15.20 with teachers (*TCHID*) being chosen as the cross section of interest. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *TCHMASTERS*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does gender or race make a difference?

b. Repeat the estimation in (a) using cluster-robust standard errors, with the cluster defined by individual teachers, *TCHID*. Are the robust standard errors larger or smaller. Compare the 95% interval estimate for the coefficient of *SMALL* using conventional and robust standard errors.

c. Reestimate the model in part (a) with teacher random effects and using both conventional and cluster-robust standard errors. Compare these results with those from parts (a) and (b).

d. Are there any variables in the equation that might be correlated with the teacher effects? Recall that teachers were randomly assigned within schools, but not across schools. Create teacher-level averages of the variables *BOY*, *WHITE_ASIAN*, and *FREELUNCH* and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

e. Suppose that we treat *FREELUNCH* as endogenous. Use the Hausman–Taylor estimator for this model. Compare the results to the OLS estimates in (a) and the random effects estimates in part (d). Do you find any substantial differences?

15.22 What is the relationship between crime and punishment? This important question has been examined by Cornwell and Trumbull[18] using a panel of data from North Carolina. The cross sections are 90 counties, and the data are annual for the years 1981–1987. The data are in the data file *crime*. In these models, the crime rate is explained by variables describing the deterrence effect of the legal system, wages in the private sector (which represents returns to legal activities), socioeconomic conditions such as population density and the percentage of young males in the population, and annual dummy variables to control for time effects. The authors argue that there may be heterogeneity across counties (unobservable county-specific characteristics).

a. What do you expect will happen to the crime rate if (i) deterrence increases, (ii) wages in the private sector increase, (iii) population density increases, and (iv) the percentage of young males increases?

b. Consider a model in which the log of crime rate (*LCRMRTE*) is a function of the log of the probability of arrest (*LPRBARR*), the log of probability of conviction (*LPRBCONV*), the log of the probability of a prison sentence (*LPRBPRIS*), the log of average prison sentence (*LAVGSEN*), and the log of average weekly wage in the manufacturing sector (*LWMFG*). Estimate this model by OLS. (i) Discuss the signs of the estimated coefficients and their significance. Are they as you expected? (ii) Interpret the coefficient on *LPRBARR*.

c. Estimate the model in (b) using a fixed effects estimator. (i) Discuss the signs of the estimated coefficients and their significance. Are they as you expected? (ii) Interpret the coefficient on *LPRBARR* and compare it to the estimate in (b). What do you conclude about the deterrent effect of the probability of arrest? (iii) Interpret the coefficient on *LAVGSEN*. What do you conclude about the severity of punishment as a deterrent?

d. In the fixed effects estimation from part (c), test whether the county level effects are all equal.

e. Based on these results, what public policies would you advocate to deal with crime in the community?

15.23 Macroeconomists are interested in factors that explain economic growth. An aggregate production function specification was studied by Duffy and Papageorgiou.[19] The data are in the data file *ces*. They consist of cross-sectional data on 82 countries for 28 years, 1960–1987.

a. Estimate a Cobb–Douglas production function

$$LY_{it} = \beta_1 + \beta_2 LK_{it} + \beta_3 LL_{it} + e_{it}$$

where *LY* is the log of GDP, *LK* is the log of capital, and *LL* is the log of labor. Interpret the coefficients on *LK* and *LL*. Test the hypothesis that there are constant returns to scale, $\beta_2 + \beta_3 = 1$.

b. Add a time trend variable $t = 1, 2, \ldots, 28$, to the specification in (a). Interpret the coefficient of this variable. Test its significance at the 5% level. What effect does this addition have on the estimates of β_2 and β_3?

c. Assume $\beta_2 + \beta_3 = 1$. Solve for β_3 and substitute this expression into the model in (b). Show that the resulting model is $LYL_{it} = \beta_1 + \beta_2 LKL_{it} + \lambda t + e_{it}$ where *LYL* is the log of the output–labor ratio, and *LKL* is the log of the capital–labor ratio. Estimate this restricted, constant returns to

[18] "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics*, 1994, 76, 360–366.

[19] "A Cross-Country Empirical Investigation of the Aggregate Production Function Specification," *Journal of Economic Growth*, 2000, 5, 83–116.

scale, version of the Cobb–Douglas production function. Compare the estimate of β_2 from this specification to that in part (b).

d. Estimate the model in (b) using a fixed effects estimator. Test the hypothesis that there are no cross-country differences. Compare the estimates to those in part (b).

e. Using the results in (d), test the hypothesis that $\beta_2 + \beta_3 = 1$. What do you conclude about constant returns to scale?

f. Estimate the restricted version of the Cobb–Douglas model in (c) using the fixed effects estimator. Compare the results to those in part (c). Which specification do you prefer? Explain your choice.

g. Using the specification in (b), replace the time trend variable t with dummy variables *D2–D28*. What is the effect of using this dummy variable specification rather than the single time trend variable?

15.24 This exercise illustrates the transformation that is necessary to produce GLS estimates for the random effects model. It utilizes the data on investment (*INV*), value (*V*) and capital (*K*) in the data file *grunfeld11*. The model is

$$INV_{it} = \beta_1 + \beta_2 V_{it} + \beta_3 K_{it} + u_i + e_{it}$$

We assume the random effects assumptions RE1–RE5 hold.

a. Find fixed effects estimates of β_2 and β_3. Check that the variance estimate that you obtain is $\hat{\sigma}_e^2 = 2530.042$.

b. Compute the sample means $\overline{INV}_i$, $\overline{V}_i$, and $\overline{K}_i$ for each of the 11 firms. [*Hint*: one way to do this to regress each of the variables (*INV*, then *V*, then *K*) on 11 indicator variables, 1 for each firm, and in each case save the predictions.]

c. Estimate β_1, β_2, and β_3 from the between regression

$$\overline{INV}_i = \beta_1 + \beta_2 \overline{V}_i + \beta_3 \overline{K}_i + u_i + \overline{e}_i.$$

Check that the variance estimate for $\sigma_*^2 = \text{var}(u_i + \overline{e}_i.)$ is $\hat{\sigma}_*^2 = 6328.554$. [*Hint*: use the predictions obtained in (b) to run the regression. If you do so, you will be using each of the N observations repeated T times. The coefficient estimates will be unaffected, but the sum of squared errors will be $T = 20$ times bigger than it should be, and the divisor used to estimate the error variance will be $NT - K$ instead of $N - K$. You will need to make adjustments accordingly.]

d. Show that

$$\hat{\alpha} = 1 - \sqrt{\frac{\hat{\sigma}_e^2}{T\hat{\sigma}_*^2}} = 0.85862$$

e. Apply least squares to the regression model

$$INV_{it}^* = \beta_1 x_1^* + \beta_2 V_{it}^* + \beta_3 K_{it}^* + v_{it}^*$$

where the transformed variables are given by $INV_{it}^* = INV_{it} - \hat{\alpha}\overline{INV}_i$, $x_1^* = 1 - \hat{\alpha}$, $V_{it}^* = V_{it} - \hat{\alpha}\overline{V}_i$, and $K_{it}^* = K_{it} - \hat{\alpha}\overline{K}_i$.

f. Use your software to obtain random effects estimates of the original equation. Compare those estimates with those you obtained in part (e).

15.25 Consider the production relationship on Chinese firms used in several chapter examples. We now add another input, *MATERIALS*. Use the data set from the data file *chemical3* for this exercise. (The data file *chemical* includes many more firms.)

$$\ln(SALES_{it}) = \beta_1 + \beta_2 \ln(CAPITAL_{it}) + \beta_3 \ln(LABOR_{it}) + \beta_4 \ln(MATERIALS_{it}) + u_i + e_{it}$$

a. Estimate this model using OLS. Compute conventional, heteroskedasticity robust, and cluster-robust standard errors. Using each type of standard error construct a 95% interval estimate for the elasticity of *SALES* with respect to *MATERIALS*. What do you observe about these intervals?

b. Using each type of standard error in part (a), test at the 5% level the null hypothesis of constant returns to scale, $\beta_2 + \beta_3 + \beta_4 = 1$ versus the alternative $\beta_2 + \beta_3 + \beta_4 \neq 1$. Are the results consistent?

c. Use the OLS residuals from (a) and carry out the $N \times R^2$ test from Chapter 9 to test for AR(1) serial correlation in the errors using the 2005 and 2006 data. Is there evidence of serial correlation? What factors might be causing it?

d. Estimate the model using random effects. How do these estimates compare to the OLS estimates? Test the null hypothesis $\beta_2 + \beta_3 + \beta_4 = 1$ versus the alternative $\beta_2 + \beta_3 + \beta_4 \neq 1$. What do you conclude. Is there evidence of unobserved heterogeneity? Carry out the LM test for the presence of random effects at the 5% level of signficance.

e. Estimate the model using fixed effects. How do the estimates compare to those in (d)? Use the Hausman test for the significance of the difference in the coefficients. Is there evidence that the unobserved heterogeneity is correlated with one or more of the explanatory variables? Explain.

f. Obtain the fixed effects residuals, $\tilde{e}_{it}$. Using OLS with cluster-robust standard errors estimate the regression $\tilde{e}_{it} = \rho \tilde{e}_{i,t-1} + r_{it}$, where r_{it} is a random error. As noted in Exercise 15.10, if the idiosyncratic errors e_{it} are uncorrelated we expect $\rho = -1/2$. Rejecting this hypothesis implies that idiosyncratic errors e_{it} are serially correlated. Using the 5% level of significance, what do you conclude?

g. Estimate the model by fixed effects using cluster-robust standard errors. How different are these standard errors from the conventional ones in part (e)?

15.26 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty per 100 students, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987. Only use data on public universities in this exercise.

a. Calculate the summary statistics for the model variables for the years 1987 and 2011. What do you observe about the sample averages of these variables?

b. Estimate the model by random effects. Discuss the signs and significance of the estimated coefficients. What is the predicted percentage cost per student change if one additional tenure track faculty is hired, per 100 students? What does the estimated value of δ_8 suggest?

c. Using the random effects estimates, test the following hypotheses at the 5% level: (i) $H_0: \beta_2 \geq \beta_3$, $H_1: \beta_2 < \beta_3$; (ii) $H_0: \beta_4 \leq \beta_6$, $H_1: \beta_4 > \beta_6$; and (iii) $H_0: \beta_5 \geq \beta_6$, $H_1: \beta_5 < \beta_6$. What do these tests imply about the relative costs of undergraduate students versus graduate students, tenure track faculty relative to contract faculty, and contract faculty relative to graduate assistants?

d. Calculate the time averages of the explanatory variables other than the indicator variables, for example, $\overline{FTESTU}_{i\bullet}$. Add these variables to the model and test their joint significance at the 1% level. What does the test result tell us about using the random effects estimator in this case? Which assumption is being tested?

e. Obtain the fixed effects estimates of the model. Discuss the signs and significance of the estimated coefficients. What is the predicted percentage cost per student change if one additional tenure track faculty is hired, per 100 students? What does the estimated value of δ_8 suggest? How do these estimates compare to the random effects estimates?

f. Using the fixed effects estimates, test the following hypotheses at the 5% level: (i) $H_0: \beta_2 \geq \beta_3$, $H_1: \beta_2 < \beta_3$; (ii) $H_0: \beta_4 \leq \beta_6$, $H_1: \beta_4 > \beta_6$; and (iii) $H_0: \beta_5 \geq \beta_6$, $H_1: \beta_5 < \beta_6$. What do these tests imply about the relative costs of undergraduate students versus graduate students, tenure track faculty relative to contract faculty, and contract faculty relative to graduate assistants?

15.27 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period from 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty per 100 students, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987.

a. Calculate the summary statistics for the model variables for the years 1987 and 2011 separately for public and private universities. What do you observe about the sample averages of these variables? In particular, what is the increase in *TC* between 1987 and 2011 for each type of university. What has happened to the number of tenure track faculty and the number of contract faculty?

b. Using OLS, estimate the model for public universities using conventional and cluster-robust standard errors. Are the standard errors noticeably different?

c. Using OLS, estimate the model for private universities using conventional and cluster-robust standard errors. Are the standard errors noticeably different? How do the coefficient estimates for the private universities compare to those for the public universities?

d. Estimate the model using fixed effects with cluster-robust standard errors for the public universities. How do these estimates compare to the OLS estimates in (b)? What are the important differences?

e. Estimate the model using fixed effects with cluster-robust standard errors for the private universities. How do these estimates compare to the estimates for the public universities in part (d)? What are the important differences?

15.28 The data file *collegecost* contains data on cost per student and related factors at four-year colleges in the U.S., covering the period 1987 to 2011. In this exercise, we explore a minimalist model predicting cost per student. Specify the model to be

$$\ln(TC_{it}) = \beta_1 + \beta_2 FTESTU_{it} + \beta_3 FTGRAD_{it} + \beta_4 TT_{it} + \beta_5 GA_{it} + \beta_6 CF_{it} + \sum_{t=2}^{8} \delta_t D_t + u_i + e_{it}$$

where *TC* is the total cost per student, *FTESTU* is number of full-time equivalent students, *FTGRAD* is number of full-time graduate students, *TT* is number of tenure track faculty per 100 students, *GA* is number of graduate assistants per 100 students, and *CF* is the number of contract faculty, which are hired on a year to year basis. The D_t are indicator variables for the years 1989, 1991, 1999, 2005, 2008, 2010, and 2011. The base year is 1987. Use data only on public universities for this question.

a. Create first differences of the variables. Using the 2011 data, estimate by OLS the first-difference model

$$\Delta\ln(TC_{it}) = \beta_2 \Delta FTESTU_{it} + \beta_3 \Delta FTGRAD_{it} + \beta_4 \Delta TT_{it} + \beta_5 \Delta GA_{it} + \beta_6 \Delta CF_{it} + \Delta e_{it}$$

b. Repeat the estimation in (a) adding an intercept term. What is the interpretation of the constant?

c. Repeat the estimation in (a) adding an intercept plus the 2011 observations on the variables *FTESTU*, *FTGRAD*, *TT*, *GA*, and *CF*. If the assumption of strict exogeneity holds none of the coefficients on these variables should be significant, and they should be jointly insignificant as well. What do you conclude? Why is this assumption important for the estimation of panel data regression models?

d. Create the one period future, or forward, value for each variable, x_{t+1}. That is, for example, in year *t* create a new variable $FTESTU_{i,t+1}$. Using data from 2008 and 2010, estimate the panel data regression model by fixed effects, including the forward values of *FTESTU*, *FTGRAD*, *TT*, *GA*, and *CF*. If the assumption of strict exogeneity holds none of the coefficients on these variables should be significant, and they should be jointly insignificant as well. What do you conclude?

15.29 In this exercise, we re-examine the data in Exercise 15.22, a panel of data from North Carolina. Consider a model in which the log of crime rate (*LCRMRTE*) is a function of the log of police per capita (*LPOLPC*), the log of the probability of arrest (*LPRBARR*), the log of the probability of conviction (*LPRBCONV*), the log of average prison sentence (*LAVGSEN*), and the log of average weekly wage in the manufacturing sector (*LWMFG*) and indicator variables for the western region (*WEST*) and urban counties (*URBAN*).

a. It is possible that the crime rate and police per capita are jointly determined and that *LPOLPC* might be endogenous. Hence we consider estimating the model by 2SLS. As instruments we use the log of tax revenue per capita (*LTAXPC*) and the log of the ratio of face-to-face crimes relative to other types of crimes (*LMIX*). Estimate the first-stage regression of *LPOLPC* on the other

variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak?

b. Using the instruments in (a), estimate the model by 2SLS. Are the deterrence variables significant?

c. Test for the endogeneity of *LPOLPC* and test the validity of the surplus instrument. What do you conclude in each case?

d. The estimation in (b) ignores unobserved county heterogeneity. For each variable, except the time-invariant variables *WEST* and *URBAN*, obtain the variables in the deviation about the county mean form, that is, apply the within transformation to each variable. Estimate the first-stage model with the variables in deviation from the mean form. Test the joint significance of the two transformed instruments.

e. Using the transformed instruments and other variables, estimate the model by 2SLS. What differences do you observe between these estimates and those in part (b)? Recall that you must adjust the standard errors for the correct degrees of freedom, as in Example 15.5. (*Note*: You may investigate whether your software has an automatic command to do 2SLS with panel data as a check.)

f. Using the transformed instruments and other variables, test for the endogeneity of *LPOLPC* and test the validity of the surplus instrument. What do you conclude in each case?

15.30 In this exercise, we extend Exercise 15.29 by also considering the possibility that the probability of arrest is jointly determined with the crime rate and the number of police per capita. The idea is that when the crime rate is high, the police may intensify their efforts to reduce crime by increasing the arrest rate. Consider the same model as in Exercise 15.29.

a. It is possible that the crime rate and police per capita are jointly determined and that *LPOLPC* and *LPRBARR* might be endogenous. Hence we consider estimating the model by 2SLS. As instruments we use the log of tax revenue per capita (*LTAXPC*) and the log of the ratio of face-to-face crimes relative to other types of crimes (*LMIX*). Estimate the first-stage regression of *LPOLPC* on the other variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak? Estimate the first-stage regression of *LPRBARR* on the other variables, except *LCRMRTE*, and the two instruments. Test the joint significance of the IV. Can we reject the null hypothesis that the IV are weak?

b. Using the instruments in (a), estimate the model, treating both *LPOLPC* and *LPRBARR* as endogenous, by 2SLS. Are the deterrence variables significant?

c. Test for the endogeneity of *LPOLPC* and *LPRBARR* using the regression-based Hausman test. What do you conclude in each case?

d. The estimation in (b) ignores unobserved county heterogeneity. For each variable, except the time-invariant variables *WEST* and *URBAN*, obtain the variables in the deviation about the county mean form, that is, apply the within transformation to each variable. Estimate the first-stage model for both *LPOLPC* and *LPRBARR* with the variables in deviation from the mean form. Test the joint significance of the two transformed instruments.

e. Using the transformed instruments and other variables, estimate the model, treating both *LPOLPC* and *LPRBARR* as endogenous, by 2SLS. What differences do you observe between these estimates and those in part (b)? Recall that you must adjust the standard errors for the correct degrees of freedom, as in Example 15.5. (*Note*: You may investigate whether your software has an automatic command to do 2SLS with panel data as a check.)

f. Test for the endogeneity of *LPOLPC* and *LPRBARR* using the regression-based Hausman test. What do you conclude in each case?

Appendix 15A Cluster-Robust Standard Errors: Some Details

To appreciate the nature of cluster-robust standard errors, we return momentarily to a simple regression model for cross-sectional data

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

Using the result $b_2 = \beta_2 + \sum_{i=1}^{N} w_i e_i$, where $w_i = (x_i - \bar{x}) \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2$, in Appendix 8A, we showed that the variance of the least squares estimator b_2, in the presence of heteroskedasticity, is given by

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N} w_i e_i|\mathbf{x}\right) = \sum_{i=1}^{N} w_i^2 \operatorname{var}(e_i|\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=i+1}^{N} 2w_i w_j \operatorname{cov}(e_i, e_j|\mathbf{x})$$

$$= \sum_{i=1}^{N} w_i^2 \operatorname{var}(e_i|\mathbf{x}) = \sum_{i=1}^{N} w_i^2 \sigma_i^2$$

Because we are assuming a random sample of cross-sectional individuals, $\operatorname{cov}(e_i, e_j|\mathbf{x}) = 0$ for $i \neq j$, leading to the simplification in the second line of the above equation.

Now suppose we have a panel simple regression model

$$y_{it} = \beta_1 + \beta_2 x_{it} + e_{it} \tag{15A.1}$$

with the assumptions $\operatorname{cov}(e_{it}, e_{is}|\mathbf{x}) = \psi_{its}$ and $\operatorname{cov}(e_{it}, e_{js}|\mathbf{x}) = 0$ for $i \neq j$. In equation (15.29) we denoted $\operatorname{var}(v_{it}) = \sigma_u^2 + \sigma_{it}^2 = \psi_{it}^2$. In this appendix we use an alternative notation, to simplify the double summations. Let $\operatorname{var}(v_{it}) = \psi_{itt} = \operatorname{cov}(v_{it}, v_{it})$. The pooled least squares estimator for β_2 is given by

$$b_2 = \beta_2 + \sum_{i=1}^{N}\sum_{t=1}^{T} w_{it} e_{it} \tag{15A.2}$$

where

$$w_{it} = \frac{x_{it} - \bar{\bar{x}}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \bar{\bar{x}}\right)^2}$$

with $\bar{\bar{x}} = \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}/NT$. The variance of the pooled least squares estimator b_2 is given by

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it} e_{it}|\mathbf{x}\right) = \operatorname{var}\left(\sum_{i=1}^{N} g_i|\mathbf{x}\right) \tag{15A.3}$$

where $g_i = \sum_{t=1}^{T} w_{it} e_{it}$ is a weighted sum of the errors for individual i. Because we have a random sample, the errors for different individuals are uncorrelated, implying that g_i is uncorrelated with g_j for $i \neq j$. Thus,

$$\operatorname{var}(b_2|\mathbf{x}) = \operatorname{var}\left(\sum_{i=1}^{N} g_i|\mathbf{x}\right) = \sum_{i=1}^{N} \operatorname{var}(g_i|\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=i+1}^{N} 2\operatorname{cov}(g_i, g_j|\mathbf{x}) = \sum_{i=1}^{N} \operatorname{var}(g_i|\mathbf{x}) \tag{15A.4}$$

To find $\operatorname{var}(g_i|\mathbf{x})$ suppose for the moment that $T = 2$, then

$$\operatorname{var}(g_i|\mathbf{x}) = \operatorname{var}\left(\sum_{t=1}^{2} w_{it} e_{it}|\mathbf{x}\right) = w_{i1}^2 \operatorname{var}(e_{i1}|\mathbf{x}) + w_{i2}^2 \operatorname{var}(e_{i2}|\mathbf{x}) + 2w_{i1}w_{i2}\operatorname{cov}(e_{i1}, e_{i2}|\mathbf{x})$$

$$= w_{i1}^2 \psi_{i11} + w_{i2}^2 \psi_{i22} + 2w_{i1}w_{i2}\psi_{i12}$$

$$= \sum_{t=1}^{2}\sum_{s=1}^{2} w_{it} w_{is} \psi_{its}$$

For $T > 2$, $\text{var}(g_i|\mathbf{x}) = \sum_{t=1}^{T}\sum_{s=1}^{T} w_{it}w_{is}\psi_{its}$. Substituting this expression into (15A.4), we have

$$\begin{aligned}\text{var}(b_2|\mathbf{x}) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} w_{it}w_{is}\psi_{its} \\ &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)\left(x_{is}-\bar{\bar{x}}\right)\psi_{its}}{\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)^2\right)^2}\end{aligned} \tag{15A.5}$$

Recall that $\text{cov}(e_{it}, e_{is}|\mathbf{x}) = E(e_{it}e_{is}|\mathbf{x}) = \psi_{its}$. A cluster-robust variance estimate is obtained from (15A.5) by replacing ψ_{its} with $\hat{e}_{it}\hat{e}_{is}$. Thus, a cluster-robust standard error for b_2 is given by the square root of

$$\widehat{\text{var}}(b_2|\mathbf{x}) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)\left(x_{is}-\bar{\bar{x}}\right)\hat{e}_{it}\hat{e}_{is}}{\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{\bar{x}}\right)^2\right)^2} \tag{15A.6}$$

The above description of how cluster-robust standard errors are calculated and the logic behind them was done in terms of a model with just one explanatory variable. To describe the robust variance estimator for models with more than one explanatory variable, matrix algebra is required, but the principle is the same.

Finally, you will find that the cluster-robust standard errors produced by most software packages apply a degrees of freedom correction to the expression in (15A.6). Unfortunately, they do not all use the same correction factor. When using a cluster-robust standard error, the effective number of observations is G, the number of clusters.[20]

Appendix 15B Estimation of Error Components

The RE model is

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) \tag{15B.1}$$

where u_i is the individual-specific error and e_{it} is the usual regression error. We will discuss the case for a balanced panel, with T time-series observations for each of N individuals. To implement GLS estimation we need to consistently estimate σ_u^2, the variance of the individual-specific error component, and σ_e^2, the variance of the regression error.

The regression error variance σ_e^2 comes from the fixed effects estimator. In (15.14), we transform the panel data regression into "**deviation about the individual mean**" form

$$y_{it} - \bar{y}_i = \beta_2(x_{2it} - \bar{x}_{2i}) + (e_{it} - \bar{e}_i) \tag{15B.2}$$

The least squares estimator of this equation yields the same estimates and sum of squared errors (denoted here by SSE_{DV}) as least squares applied to a model that includes a dummy variable for each individual in the sample. A consistent estimator of σ_e^2 is obtained by dividing SSE_{DV} by the

[20] See Carter, et al. "Asymptotic Behavior of a *t*-Test Robust to Cluster Heterogeneity," *The Review of Economics and Statistics*, 2017, 99(4), 698–709.

appropriate degrees of freedom, which is $NT - N - K_S$, where K_S is the number of parameters that are present in the transformed model (15B.2)

$$\hat{\sigma}_e^2 = \frac{SSE_{DV}}{NT - N - K_S} \tag{15B.3}$$

The estimator of σ_u^2 requires a bit more work. We begin with the time-averaged observations in (15.13)

$$\bar{y}_i = \beta_1 + \beta_2 \bar{x}_{2i} + \alpha_1 w_{1i} + u_i + \bar{e}_i, \quad i = 1, 2, \ldots, N \tag{15B.4}$$

The least squares estimator of (15B.4) is called the **between estimator**, as it uses variation between individuals as a basis for estimating the regression parameters. This estimator is unbiased and consistent, but not minimum variance under the error assumptions of the random effects model. The error term in this model is $u_i + \bar{e}_i$; it is uncorrelated across individuals, and has homoskedastic variance

$$\begin{aligned} \operatorname{var}(u_i + \bar{e}_i) &= \operatorname{var}(u_i) + \operatorname{var}(\bar{e}_i) = \operatorname{var}(u_i) + \operatorname{var}\left(\sum_{t=1}^{T} e_{it}/T\right) \\ &= \sigma_u^2 + \frac{1}{T^2}\operatorname{var}\left(\sum_{t=1}^{T} e_{it}\right) = \sigma_u^2 + \frac{T\sigma_e^2}{T^2} \\ &= \sigma_u^2 + \frac{\sigma_e^2}{T} \end{aligned} \tag{15B.5}$$

We can estimate the variance in (15B.5) by estimating the between regression in (15B.4), and dividing the sum of squared errors, SSE_{BE}, by the degrees of freedom $N - K_{BE}$, where K_{BE} is the total number of parameters in the between regression, including the intercept parameter. Then

$$\widehat{\sigma_u^2 + \frac{\sigma_e^2}{T}} = \frac{SSE_{BE}}{N - K_{BE}} \tag{15B.6}$$

With this estimate in hand, we can estimate σ_u^2 as

$$\hat{\sigma}_u^2 = \widehat{\sigma_u^2 + \frac{\sigma_e^2}{T}} - \frac{\hat{\sigma}_e^2}{T} = \frac{SSE_{BE}}{N - K_{BE}} - \frac{SSE_{DV}}{T(NT - N - K_S)} \tag{15B.7}$$

We have obtained the estimates of σ_u^2 and σ_e^2 using what is called the Swamy–Arora method. This method is implemented in software packages and is well established. We note, however, that it is possible in finite samples to obtain an estimate $\hat{\sigma}_u^2$ in (15B.7) that is negative, which is obviously infeasible. If this should happen, one option is simply to set $\hat{\sigma}_u^2 = 0$, which implies that there are no random effects. Alternatively, your software may offer other options for estimating the variance components, which you might try.

CHAPTER 16

Qualitative and Limited Dependent Variable Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Give some examples of economic decisions in which the observed outcome is a binary variable.
2. Explain why probit, or logit, is usually preferred to least squares when estimating a model in which the dependent variable is binary.
3. Give some examples of economic decisions in which the observed outcome is a choice among several alternatives, both ordered and unordered.
4. Compare and contrast the multinomial logit model to the conditional logit model.
5. Give some examples of models in which the dependent variable is a count variable.
6. Discuss the implications of censored data for least squares estimation.
7. Describe what is meant by the phrase ‘‘sample selection.’’

KEYWORDS

alternative specific variables
binary choice models
censored data
conditional logit
count data models
feasible generalized least squares
Heckit
identification problem
independence of irrelevant alternatives (IIA)
index models
individual specific variables
latent variables
likelihood function
likelihood ratio
limited dependent variables
linear probability model
logistic random variable
logit
log-likelihood function
marginal effect
maximum likelihood estimation
multinomial choice models
multinomial logit
ordered probit
ordinal variables
Poisson random variable
Poisson regression model
probability ratio
probit
selection bias
Tobit model
truncated regression

In this book, we focus primarily on econometric models in which the dependent variable is continuous and fully observable; quantities, prices, and outputs are examples of such variables. However, microeconomics is a general theory of choice, and many of the choices that individuals and firms make cannot be measured by a continuous outcome variable. In this chapter, we examine some fascinating models that are used to describe choice behavior, and which do not have the usual continuous dependent variable. Our descriptions will be brief, since we will not go into all the theory, but we will reveal to you a rich area of economic applications.

We also introduce a class of models with dependent variables that are *limited*. By that we mean that they are continuous but that their range of values is constrained in some way, and their values not completely observable. Alternatives to least squares estimation must be considered for such cases, since the least squares estimator is both biased and inconsistent.

16.1 Introducing Models with Binary Dependent Variables

Many of the choices that individuals and firms make are "either-or" in nature. For example, a high-school graduate decides either to attend college or not. A worker decides either to drive to work or to get there using a different means of transportation. A household decides either to purchase a house or to rent. A firm decides either to advertise its product in a local newspaper or it decides not to. As economists we are interested in explaining why particular choices are made, and what factors enter into the decision process. We also want to know *how much* each factor affects the outcome and how to predict outcomes. Such questions lead us to the problem of constructing a statistical model of binary, either-or, choices. Such choices can be represented by a binary (indicator) variable that takes the value 1 if one outcome is chosen and the value 0 otherwise. The binary variable describing a choice is the dependent variable rather than an independent variable. This fact affects our choice of a statistical model.

The list of economic applications in which choice models may be useful is a long one. These models are useful in any economic setting in which an agent must choose one of two alternatives. Examples include the following:

- An economic model explaining why some individuals take a second or third job and engage in "moonlighting."
- An economic model of why some legislators in the U.S. House of Representatives vote for a particular bill and others do not.
- An economic model explaining why some loan applications are accepted and others are not at a large metropolitan bank.
- An economic model explaining why some individuals vote for increased spending in a school board election and others vote against.
- An economic model explaining why some female college students decide to study engineering and others do not.

This list illustrates the great variety of circumstances in which a model of binary choice may be used. In each case, an economic decision-maker chooses between two mutually exclusive outcomes.

The key feature of **binary choice models** is the nature of the outcome variable. It is an indicator variable representing the choice between two alternatives. We represent the ith individual's choice as

$$y_i = \begin{cases} 1 & \text{alternative one is chosen} \\ 0 & \text{alternative two is chosen} \end{cases} \tag{16.1}$$

Individuals make choices to maximize their utility, or well-being, and we economists would like to understand the process. What are the important factors leading to the choice and how much weight is given to each? Can we predict what the choice will be? These questions lead us to

consider how individuals make their decisions, how to build an econometric model of the choice process, and how to model the probability of choosing one alternative or the other.

It's always best to start at the beginning. Unlike the outcome of a game of chance, such as flipping a coin and observing a head or a tail, the probability that alternative one will be chosen varies from individual to individual, and the probability depends on many factors, describing the individual and the characteristics of the alternatives. As in a regression model, let these factors be denoted $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. Then the conditional probability that the ith individual chooses alternative one is $P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i)$, where $p(\mathbf{x}_i)$ is a function of the factors $\mathbf{x}_i$, and because it is a probability, $0 \le p(\mathbf{x}_i) \le 1$. The conditional probability of choosing alternative two is $P(y_i = 0|\mathbf{x}_i) = 1 - p(\mathbf{x}_i)$. We can represent the conditional probability function for the random variable y_i in equation (16.1) as

$$f(y_i|\mathbf{x}_i) = p(\mathbf{x}_i)^{y_i}\left[1 - p(\mathbf{x}_i)\right]^{1-y_i} \qquad y_i = 0, 1 \tag{16.2}$$

Then $P(y_i = 1|\mathbf{x}_i) = f(1|\mathbf{x}_i) = p(\mathbf{x}_i)$ and $P(y_i = 0|\mathbf{x}_i) = f(0|\mathbf{x}_i) = 1 - p(\mathbf{x}_i)$. The standard models of probabilistic choice are simply alternative ways of representing, or approximating, $P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i)$.

EXAMPLE 16.1 | A Transportation Problem

An important problem in transportation economics is explaining an individual's choice between driving (private transportation) and taking the bus (public transportation) when commuting to work, assuming, for simplicity, that these are the only two alternatives. We can imagine many factors that affect the choice, including an individual's characteristics, such as age, income, and sex; the characteristics of their automobile, such as its reliability, comfort, and fuel economy; the characteristics of the public transportation, such as reliability, cost, and safety. In our example, we will focus on a single factor, commuting time. Define the explanatory variable

x_i = (commuting time by bus
− commuting time by car, for the ith individual)

A priori we expect that as x_i increases, and commuting time by bus increases relative to commuting time by car, and holding all else constant, an individual would be more inclined to drive. Suppose that alternative one is driving to work, $y_i = 1$, and alternative two is taking public transportation, $y_i = 0$. Then the probability that the ith individual drives to work is $P(y_i = 1|x_i) = p(x_i)$. Our reasoning suggests that there is a positive relationship between the difference in commuting time and the probability that an individual will drive to work. Using data on individuals and their choices, we will obtain estimates of how much increases in commuting time by bus relative to driving will affect the probability that an individual will drive. Using the estimates, we can predict the choice of an individual when the commuting time by bus is, for example, 20 minutes longer than the commuting time by car. We will also develop methods for testing hypotheses about the nature of the relationship, such as testing whether the difference in commuting time is a statistically significant factor in the decision.

16.1.1 The Linear Probability Model

We discussed the **linear probability model** in Sections 7.4 and 8.7. It is a regression model that arises straightforwardly from the definition of expected value. Using the probability model in (16.2),

$$E(y_i|\mathbf{x}_i) = \sum_{y_i=0}^{1} y_i f(y_i|\mathbf{x}_i) = 0 \times f(0|\mathbf{x}_i) + 1 \times f(1|\mathbf{x}_i) = p(\mathbf{x}_i) \tag{16.3}$$

The population average outcome, the average choice, is the probability that the first alternative is chosen. It is natural to specify a linear regression model for the probability

$$p(\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \tag{16.4}$$

Let the random error e_i account for the difference between the observed outcome y_i and the conditional mean $E(y_i|\mathbf{x}_i)$,

$$e_i = y_i - E(y_i|\mathbf{x}_i) \tag{16.5}$$

Then

$$y_i = E(y_i|\mathbf{x}_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{16.6}$$

If $E(e_i|\mathbf{x}_i) = 0$, then the least squares estimator of the parameters is unbiased, or if random error e_i is uncorrelated with $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$, then the least squares estimator is consistent. These are the usual OLS properties.

For a continuous variable x_{ik}, the **marginal effect** is

$$\partial E(y_i|\mathbf{x}_i)/\partial x_{ik} = \beta_k \tag{16.7}$$

Here is where some difficulty enters. Suppose that $\beta_k > 0$. Increasing x_{ik} by one unit increases $p(\mathbf{x}_i)$, the probability of alternative one being chosen, by a constant amount β_k. This puts us into the uncomfortable position of concluding that the probability can become one, or greater than one, if x_{ik} becomes large enough. Similarly, if $\beta_k < 0$ then the probability of alternative one being chosen can become negative if x_{ik} becomes large enough. These are the logical inconsistencies in the linear probability model. It is because of these difficulties that we develop alternatives to the linear probability model in Section 16.2. Nevertheless, the regression model approach is very familiar, and by now easy, and it is a useful approximation tool for the purpose of estimating marginal effects in nonextreme cases.

Apart from the logical problem noted above, which is important, there are two other more minor consequences of using the linear probability model. First, since y_i takes only two values, one and zero, it must be true that $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ takes the same two values. If $y_i = 1$, then it follows that $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i = 1$, so that

$$e_i = 1 - (\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$$

If $y_i = 0$, then $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i = 0$ so that

$$e_i = -(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$$

This seems very odd—the random error that accounts for all omitted factors and other specification errors takes only two values. This is the result of imposing a linear regression structure on a choice problem in which the outcome is binary, one or zero.

Secondly, the conditional variance in the random error is

$$\text{var}(e_i|\mathbf{x}_i) = p(\mathbf{x}_i)\left[1 - p(\mathbf{x}_i)\right] = \sigma_i^2 \tag{16.8}$$

and is necessarily heteroskedastic. When estimating the linear probability model, this feature must be recognized. When using the OLS estimator, we must at least use heteroskedasticity robust standard errors. Alternatively use the FGLS, **feasible generalized least squares**, estimation methodology discussed in Section 8.6.

EXAMPLE 16.2 | A Transportation Problem: The Linear Probability Model

Ben-Akiva and Lerman[1] have sample data on automobile and public transportation travel times and the alternative chosen for $N = 21$ individuals in the data file *transport*. The variable *AUTO* is an indicator variable taking the value one if automobile transportation is chosen and is zero if public transportation is chosen,

$$AUTO = \begin{cases} 1 & \text{auto is chosen} \\ 0 & \text{public transportation (bus) is chosen} \end{cases}$$

[1] (1985) *Discrete Choice Analysis*, MIT Press.

The variables *AUTOTIME* and *BUSTIME* are minutes of commuting time. The explanatory variable we consider is $DTIME = (BUSTIME - AUTOTIME) \div 10$, which is the commuting time differential in 10-minute increments. The linear probability model is $AUTO_i = \beta_1 + \beta_2 DTIME_i + e_i$. The OLS fitted model, with heteroskedasticity robust standard errors, is

$$\widehat{AUTO}_i = 0.4848 + 0.0703 DTIME_i \quad R^2 = 0.61$$
$$\text{(robse)} \quad (0.0712) \quad (0.0085)$$

We estimate that if travel times by public transportation and automobile are equal, so that $DTIME = 0$, then the probability of a person choosing automobile travel is 0.4848, close to 50–50, with a 95% interval estimate of [0.34, 0.63]. We estimate that, holding all else constant, an increase of 10 minutes in the difference in travel time, increasing public transportation travel time relative to automobile travel time, increases the probability of choosing automobile travel by 0.07, with a 95% interval estimate of [0.0525, 0.0881], which seems relatively precise. In truth, any judgment about precision depends on the use to which the results will be put. The fitted model can be used to estimate the probability of automobile travel for any commuting time differential. For example, if $DTIME = 1$, a 10-minute longer commute by public transportation, we estimate the probability of automobile travel to be $\widehat{AUTO}_i = 0.4848 + 0.0703(1) = 0.5551$.

How well does the model fit the data? The $R^2 = 0.61$ suggests that 61% of the variation in the outcome variable is explained by the model. With probability models, we can examine how well the model predicts the outcomes. Let's predict the choice using a probability threshold of 0.50. That is, if $\widehat{AUTO}_i \geq 0.50$ we predict that a person will drive to work, and otherwise, we predict that a person will use public transportation. In the sample of 21 individuals, 10 drove to work and 11 used public transportation. Using the classification rule, we successfully predict 9 of the 10 drivers, and 10 of the 11 bus riders. That is 19 successful predictions out of the 21 cases. Looking at individual estimated probabilities of driving, we find three negative values. If the commute is 69 minutes or less by public transportation, then the estimated probability of driving is zero or negative. If commuting time is 73 minutes or more by public transportation, then the estimated probability of driving is one or greater.

16.2 Modeling Binary Choices

It is the probability of choosing one alternative or the other that is the key concept when modeling binary choice. Probabilities must be between zero and one, and the flaw in the linear probability model in Section 16.1 is that it does not impose this constraint. We now turn to two nonlinear models for binary choices, the **probit model** and the **logit model**, which ensure that choice probabilities remain between zero and one. To keep the choice probability $p(\mathbf{x}_i)$ within the interval (0, 1), a nonlinear S-shaped "sigmoid" curve can be used. In Figure 16.1(a), one such curve is illustrated for the case of a single explanatory variable, x. If, for example, $\beta_2 > 0$, then, as x increases, and, $\beta_1 + \beta_2 x$ increases, the probability curve rises rapidly at first, and then begins to increase at a decreasing rate, keeping the probability less than one no matter how large x becomes. In the other direction, the probability approaches but never reaches zero. The *slope* of the probability curve, $dp(x_i)/dx$, is the change in probability given a unit change in x. It is the **marginal effect** and, unlike in the linear probability model, the slope is not constant.

The curve shown in Figure 16.1(a) is the cumulative distribution function (*cdf*) of the standard normal random variable. This choice of the S-curve leads to a model called **probit**. Any *cdf* function for a continuous random variable will work, and many have been tried over the years. These days the main competitor to the standard normal *cdf* is the *cdf* of a **logistic random variable**, leading to a model called **logit**. In binary choice cases, probit and logit provide very similar inferences. Economists tend to choose probit rather than logit in individual choice applications because it follows logically from utility maximizing behavior and random utility models (RUMs) under the assumption that the unobserved components of utility for the two alternatives are jointly normal. To obtain a logit model within this framework, the unobserved components of utility for the two alternatives must be statistically independent and have an unusual probability density function (*pdf*).[2] However, the logit model is widely used in many disciplines and leads to very convenient generalizations. We will discuss both the probit and logit models.

[2]For more on RUM and choice models, see Appendix 16B. Also Kenneth Train (2009) *Discrete Choice Methods with Simulation, Second Edition*, Cambridge University Press.

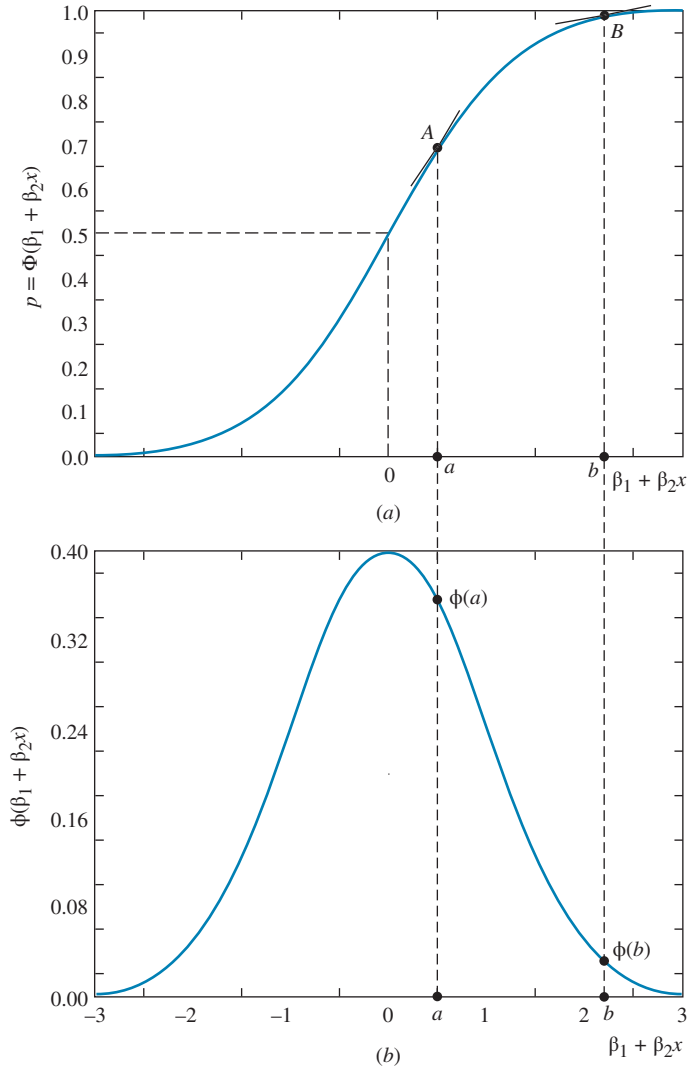


FIGURE 16.1 **(*a*) Standard normal *cdf*; (*b*) standard normal *pdf*.**

16.2.1 The Probit Model for Binary Choice

As noted above, the probit model is based on the standard normal *cdf*. If Z is a standard normal random variable, then its *pdf* is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} \quad -\infty < z < \infty \tag{16.9a}$$

The *cdf* of the standard normal distribution is

$$\Phi(z) = P[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0.5u^2} du \tag{16.9b}$$

This integral expression is the probability that a standard normal random variable falls to the left of point z. In geometric terms, it is the area under the standard normal *pdf* to the left of z. The function $\Phi(z)$ is the *cdf* that we have worked with to compute normal probabilities.

The probit statistical model expresses the probability $p(\mathbf{x}_i)$ that alternative one is chosen, $y_i = 1$, to be

$$P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i) = P[Z \le \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}] = \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}) \quad (16.10)$$

where $\Phi(z)$ is the standard normal *cdf*. The probit model is said to be *nonlinear* because (16.10) is a nonlinear function of the parameters $\beta_1, \ldots, \beta_K$. If the parameters $\beta_1, \ldots, \beta_K$ were known, we could use (16.10) to find the probability that alternative one is chosen for any set of predictor values $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. Because these parameters are not known we will estimate them.

16.2.2 Interpreting the Probit Model

Interpreting the probit model requires a bit of work. How we proceed to measure the impact of any one variable x_{ik} depends on whether it is continuous or discrete, like an indicator variable. When an explanatory variable is continuous, we can examine the marginal effect of a change in its value on the probability $p(\mathbf{x}_i)$. When the explanatory variable is an indicator variable, we can calculate the difference in the probability $p(\mathbf{x}_i)$ associated with $x_{ik} = 0$ and $x_{ik} = 1$. In both of these cases, we must deal with the fact that the magnitudes of the effects depend not only on the parameter values, $\beta_1, \ldots, \beta_K$, but also on the values of the explanatory variables, $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. We will examine these cases separately.

Marginal Effect of a Continuous Explanatory Variable

If x_k is a continuous variable then we can calculate the marginal effect by finding the derivative of (16.10). The marginal effect is

$$\frac{\partial p(\mathbf{x}_i)}{\partial x_{ik}} = \frac{\partial \Phi(t_i)}{\partial t_i} \cdot \frac{\partial t_i}{\partial x_{ik}} = \phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\beta_k \quad (16.11)$$

where $t_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$ and $\phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$ is the standard normal *pdf* evaluated at $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$. To obtain this result, we have used the chain rule of differentiation (see Derivative Rule 9 in Appendix A.3.1). Note that the marginal effect includes the *pdf* of the standard normal random variable, $\phi(\bullet)$.

To simplify the algebra, suppose that there is a single continuous explanatory variable, x. Then, the probit probability model is $p(x_i) = P[Z \le \beta_1 + \beta_2 x_i] = \Phi(\beta_1 + \beta_2 x_i)$. Assuming $\beta_2 > 0$, this is the equation of the sigmoid S-shaped curve in Figure 16.1(a). At point A in Figure 16.1(a), where $\beta_1 + \beta_2 x_i = a$, the marginal effect of a change in x on the probability is the slope of the tangent line. At point B in Figure 16.1(a), where $\beta_1 + \beta_2 x_i = b$ and the probability $\Phi(b)$ is larger, the marginal effect is smaller, which it must be to keep the probability function less than one as x increases.

The equation of the marginal effect $dp(x_i)/dx_i = \phi(\beta_1 + \beta_2 x_i)\beta_2$ is the slope of the probability function at the point $\beta_1 + \beta_2 x_i$. The *pdf* $\phi(\beta_1 + \beta_2 x_i)$, plotted in Figure 16.1(b), appears in the marginal effect because of its relationship to the cumulative distribution function $\Phi(\beta_1 + \beta_2 x_i)$. As noted in (16.9), the *cdf* is the integral of the *pdf*, and it follows that the *pdf* is the derivative of the *cdf* in (16.11). The marginal effect at point A is larger because $\phi(a) > \phi(b)$. The marginal effect equation, $dp(x_i)/dx_i = \phi(\beta_1 + \beta_2 x_i)\beta_2$, has the following implications.

1. Since $\phi(\beta_1 + \beta_2 x_i)$ is a *pdf* its value is always *positive*. Consequently, the sign of $dp(x_i)/dx_i$ is determined by the sign of β_2. If $\beta_2 > 0$ then $dp(x_i)/dx_i > 0$, and if $\beta_2 < 0$ then $dp(x_i)/dx_i < 0$.
2. As x_i changes the value of the function $\phi(\beta_1 + \beta_2 x_i)$ changes. The standard normal *pdf* reaches its maximum when $\beta_1 + \beta_2 x_i = 0$. In this case $p(x_i) = P[Z \le 0] = \Phi(0) = 0.5$; the alternatives one and two are equally likely to be chosen. It makes sense that in this case the effect of a change in x_i has its greatest effect, the marginal effect is largest, because the individual is "on the borderline."

3. On the other hand, if $\beta_1 + \beta_2 x_i$ is large, say near 3, then the probability that the individual chooses alternative one, $p(x_i)$, is very large and close to 1. In this case, a change in x_i will have relatively little effect since $\phi(\beta_1 + \beta_2 x_i)$ is nearly 0. The same is true if $\beta_1 + \beta_2 x_i$ is a large negative value, say near -3. These results are consistent with the notion that if an individual is "set" in their ways, with $p(x_i)$ near 0 or 1, the effect of a small change in x_i will be negligible.

Discrete Change Effect of an Indicator Explanatory Variable The marginal effect in (16.11) is valid only if the explanatory variable x_k is continuous. If x_k is a discrete variable, such as an indicator variable for an individual's sex, then the derivative in (16.11) cannot be used. Instead we can compute the discrete change in probability effect of x_k changing from zero to one,

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i | x_{ki} = 1) - p(\mathbf{x}_i | x_{ki} = 0) \tag{16.12a}$$

To simplify the notation, suppose $p(\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \delta D_i)$ where D_i is an indicator variable. The difference in the probability of choosing alternative one given $D_i = 1$ as compared to when $D_i = 0$ is

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i | D_i = 1) - p(\mathbf{x}_i | D_i = 0) = \Phi(\beta_1 + \beta_2 x_{i2} + \delta) - \Phi(\beta_1 + \beta_2 x_{i2}) \tag{16.12b}$$

The change can be positive or negative, depending on the sign of the parameter δ. If $\delta > 0$, then there is an increase in the probability of choosing alternative one. If $\delta < 0$, then the probability of choosing alternative one decreases. Note that the magnitude of the effect depends on the sign and magnitude of the parameter δ but also on the values of the other explanatory variables and their parameters.

Discrete Change Effect of any Explanatory Variable The use of the discrete change approach is not limited to indicator variables. It can also be used for an explanatory variable that is a count, such as $x_3 = 0, 1, 2, \ldots$ Suppose that y_i is an individual's health outcome, such as whether their blood pressure reading is too high, or not, and x_3 is the person's number of periods of exercise per week. We might be interested in the change in the probability of high blood pressure of increasing from one workout per week to three workouts per week. The discrete change approach can also be used for a continuous variable. Suppose that x_3 is the number of minutes of exercise per week. We might be interested in the change in the probability of high blood pressure of increasing the number of minutes of exercise from 90 to 120 per week. In general, suppose that we are interested in the change $x_{i3} = c$ to $x_{i3} = c + \delta$. Then the discrete change in probability is

$$\begin{aligned} \Delta p(\mathbf{x}_i) &= p(\mathbf{x}_i | x_{i3} = c + \delta) - p(\mathbf{x}_i | x_{i3} = c) \\ &= \Phi(\beta_1 + \beta_2 x_{i2} + \beta_3 c + \beta_3 \delta) - \Phi(\beta_1 + \beta_2 x_{i2} + \beta_3 c) \end{aligned} \tag{16.12c}$$

Because the model is nonlinear, the values of c and δ will affect the change in probability.

Estimating Marginal and Discrete Change Effects In order to estimate the marginal effect in (16.11) or the discrete change effect (16.12), we must have parameter estimates, $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$. The estimates are obtained by **maximum likelihood estimation**, which we will discuss in Section 16.2.3. For the moment, suppose that we have these estimates. In practice, they are obtained just like OLS estimates, with a simple computer command. Focus now on the possible values of the explanatory variables $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. There are several options for reporting marginal effects:

1. **Marginal effect at means (MEM)**[3] One choice is $\bar{\mathbf{x}} = \left(1, \bar{x}_2, \ldots, \bar{x}_K\right)$ where $\bar{x}_k$ is the sample mean of the values for the kth explanatory variable. There are two points of interest here. First, unlike the linear regression model, the fitted probit model does not pass through the "point of the means," so choosing the point $\bar{\mathbf{x}}$ has no special significance. Second, for an indicator variable, such as $x_{ik} = 1$ for females and $x_{ik} = 0$ for males, the average value $\bar{x}_k$ is the fraction of the sample that is female. Instead of a 1 or a 0, we might have $\bar{x}_k = 0.53$, indicating that 53% of the sample is female.
2. **Marginal effect at a representative value (MER)** Another possibility is to choose the values of $\mathbf{x}_i = \left(x_{i1} = 1, x_{i2}, \ldots, x_{iK}\right)$ to reflect a particular scenario, a set of values that tell a "story" about the results. That is, suppose that x_{i2} is a person's years of schooling, x_{i3} is the person's sex (1 = female), and x_{i4} is their income ($1000s). We might specify $\mathbf{x}_i = \left(1, x_{i2} = 14, x_{i3} = 1, x_{i4} = 100\right)$, representing a female with 14 years of schooling and $100,000 income. This approach is more work because the representative values for the variables should have some meaning within the context of the research problem, but in some sense, it is also the most meaningful when describing the results. Of course, some of the variables' representative values might be variable means, medians, or quartiles.
3. **Average marginal effect (AME)** A third option is to calculate the sample average marginal effect. For a continuous variable, the AME is the sample average of (16.11) evaluated at each sample observation,

$$\text{AME}\left(x_k\right) = N^{-1}\sum_{i=1}^{N}\partial p\left(\mathbf{x}_i\right)/\partial x_{ik} = \beta_k\sum_{i=1}^{N}\phi\left(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}\right)/N \tag{16.13a}$$

   For a discrete variable, we average the differences in (16.12a). In the simple model $p\left(\mathbf{x}_i\right) = \Phi\left(\beta_1 + \beta_2 x_{i2} + \delta D_i\right)$, this average is

$$\begin{aligned}\text{AME}(D) &= N^{-1}\sum_{i=1}^{N}\Delta p\left(\mathbf{x}_i\right)\\ &= \sum_{i=1}^{N}\Phi\left(\beta_1 + \beta_2 x_{i2} + \delta\right)/N - \sum_{i=1}^{N}\Phi\left(\beta_1 + \beta_2 x_{i2}\right)/N\end{aligned} \tag{16.13b}$$

   If, for example, $D_i = 1$ if a person is female, then the first term $\sum_{i=1}^{N}\Phi\left(\beta_1 + \beta_2 x_{i2} + \delta\right)/N$ assigns the female sex to everyone in the sample, and the second term $\sum_{i=1}^{N}\Phi\left(\beta_1 + \beta_2 x_{i2}\right)/N$ assigns the male sex to everyone in the sample. There are two advantages to computing the AME. First, it relieves us of having to make a choice about what to do. Second, relying on a "law of large numbers" argument, the sample average marginal, or discrete change, effect can be thought of as estimating the population average response to a change in a variable.
4. **A Histogram** A fourth option is to examine a histogram of the marginal effects computed for each $\mathbf{x}_i$ in the sample.

Predicting Choice with a Probit Model Last but not least, we can use the probit model to not only estimate the probability that an individual chooses one alternative or another but also predict the choice they will make. The probability model is $p\left(\mathbf{x}_i\right) = \Phi\left(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}\right)$. Given values of the explanatory variables, and parameter estimates, $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$, we can estimate the probability that an individual will choose alternative one as $\tilde{p}\left(\mathbf{x}_i\right) = \Phi\left(\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + \cdots + \tilde{\beta}_K x_{iK}\right)$. By comparing the estimated probability to a suitable threshold, τ, we can predict choice. The first threshold that comes to mind is 0.5. If we estimate the probability to be greater than or equal to 0.5 we predict $\tilde{y}_i = 1$, and if the estimated probability is less than 0.5, then we predict $\tilde{y}_i = 0$.

The threshold 0.5 is not necessarily the best threshold value to use. For example, suppose that we are the loan officer at a lending institution and must decide whether to give a loan to an

[3] We use the abbreviations MEM, MER, and AME, following Cameron and Trivedi (2010) *Microeconometrics Using Stata, Second Edition*, pp. 343–356.

applicant. Using data on previous borrowers, we can estimate a probit model for whether a loan was repaid on time, $y_i = 1$, or not, $y_i = 0$, as a function of borrower and loan characteristics. The fact is that most borrowers do repay their loans. If 90% of borrowers pay their loan back and if our applicant's estimated probability of repayment is 0.60, then that is a weak endorsement for giving a loan. For a lender, choosing the profit maximizing threshold τ^* is not an easy task. The correct decision is to give a loan to someone who will repay it and to not give a loan to someone who won't repay it. Lenders must weigh two types of incorrect decisions. If the lender gives a loan to someone who does not repay it, then there are costs (losses) associated with collecting the loan; further correspondence, legal action, and so on. If the lender does not give a loan to someone who would repay, there are foregone profits, opportunity costs. Lenders must compare the costs of these errors. If the threshold is raised, there are increased foregone profits; if the threshold is lowered, there are more collection costs. There is no one universal threshold that is suitable for every type of situation.

16.2.3 Maximum Likelihood Estimation of the Probit Model

The maximum likelihood estimation (MLE) methodology is discussed in Appendix C.8. Maximum likelihood estimation is based on a principle that is an alternative to the least squares principle or to other principles such as generalized least squares, or the method of moments, although it sometimes yields the same results. The MLE methodology is well suited to models we discuss in this chapter, including the probit binary choice model. Under some suitable conditions, maximum likelihood estimators have properties that are valid in large samples. If $\tilde{\beta}_k$ is the maximum likelihood estimator of the parameter β_k, then it is a consistent estimator, plim $\tilde{\beta}_k = \beta_k$, and it has an approximate normal distribution in large samples, $\tilde{\beta}_k \overset{a}{\sim} N\left[\beta_k, \text{var}(\tilde{\beta}_k)\right]$. The estimator variance is known (though complicated algebraically) and can be consistently estimated in several ways. If $\widehat{\text{var}}(\tilde{\beta}_k)$ is a consistent estimator of $\text{var}(\tilde{\beta}_k)$, then we can calculate a standard error, $\text{se}(\tilde{\beta}_k) = \sqrt{\widehat{\text{var}}(\tilde{\beta}_k)}$. Using the standard error, we can compute interval estimates, $\tilde{\beta}_k \pm z_{(1-\alpha/2)}\text{se}(\tilde{\beta}_k)$, carry out "$t$-tests," and so on in the usual way. All of these theoretical results are illustrated in Appendix C.8. In Example 16.3, we present the essence of the maximum likelihood estimation method.

EXAMPLE 16.3 | Probit Maximum Likelihood: A Small Example

We first illustrate the idea of maximum likelihood estimation in an abbreviated version of the transportation choice model from Examples 16.1 and 16.2. Suppose that we randomly select three individuals and observe that the first two drive to work and the third takes the bus; $y_1 = 1$, $y_2 = 1$, $y_3 = 0$. Furthermore, suppose that the differences in commuting times for these individuals, in 10-minute units, are $x_1 = 1.5$, $x_2 = 0.6$, $x_3 = 0.7$. What is the joint probability of observing $y_1 = 1$, $y_2 = 1$, $y_3 = 0$? The probability function for y_i is given by (16.2), which we now combine with the probit model (16.10) to obtain

$$f(y_i|x_i) = \left[\Phi(\beta_1 + \beta_2 x_i)\right]^{y_i}\left[1 - \Phi(\beta_1 + \beta_2 x_i)\right]^{1-y_i}, \quad y_i = 0, 1$$

If the three individuals are independently drawn, then the joint *pdf* for y_1, y_2, and y_3 is the product of the marginal probability functions:

$$f(y_1, y_2, y_3|x_1, x_2, x_3) = f(y_1|x_1)\, f(y_2|x_2)\, f(y_3|x_3)$$

Consequently, the probability of observing $y_1 = 1$, $y_2 = 1$, and $y_3 = 0$ is

$$P(y_1 = 1, y_2 = 1, y_3 = 0|x_1, x_2, x_3) = f(1, 1, 0|x_1, x_2, x_3) = f(1|x_1)\, f(1|x_2)\, f(0|x_3)$$

Substituting the y and x values, we have

$$\begin{aligned} P(y_1 = 1, y_2 = 1, y_3 = 0|x_1, x_2, x_3) &= \Phi\left[\beta_1 + \beta_2(1.5)\right] \times \Phi\left[\beta_1 + \beta_2(0.6)\right] \\ &\quad \times \left\{1 - \Phi\left[\beta_1 + \beta_2(0.7)\right]\right\} \\ &= L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x}) \end{aligned} \tag{16.14}$$

In statistics, the function (16.14), which gives us the probability of observing the sample data, is called the **likelihood function**. The notation $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ indicates that the likelihood function is a function of the unknown parameters once we are given the data. It is intuitively reasonable to use as estimates those values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ that maximize the probability,

or likelihood, of the observed outcome. Unfortunately, for the probit model, there are no formulas that give us the values for $\tilde{\beta}_1$ and $\tilde{\beta}_2$ as there are in least squares estimation of the linear regression model. Consequently, we must use the computer and techniques from numerical analysis to find the values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ that maximize $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$. In practice, instead of maximizing (16.14), we maximize the logarithm of (16.14), which is called the **log-likelihood function**

$$\begin{aligned}\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x}) &= \ln\left\{\Phi[\beta_1 + \beta_2(1.5)] \times \Phi[\beta_1 + \beta_2(0.6)]\right. \\ &\quad \left.\times \left\{1 - \Phi[\beta_1 + \beta_2(0.7)]\right\}\right\} \\ &= \ln \Phi[\beta_1 + \beta_2(1.5)] + \ln\Phi[\beta_1 + \beta_2(0.6)] \\ &\quad + \ln\left\{1 - \Phi[\beta_1 + \beta_2(0.7)]\right\} \qquad (16.15)\end{aligned}$$

On the surface, this appears to be a difficult task, because $\Phi(z)$ from (16.9) is such a complicated function. As it turns out, however, using a computer to maximize (16.15) is a relatively easy process.

The maximization of the log-likelihood function $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ is easier than the maximization of (16.14), because it is a sum of terms and not a product of terms. The logarithm is a nondecreasing, or monotonic, function so that the maximum values of the two functions $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ and $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ occur at the same values of β_1 and β_2, namely, $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The value of the log-likelihood function (16.15) evaluated at the maximizing values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ is very useful for hypothesis testing, which is discussed in Sections 16.2.4 and 16.2.5. Using econometric software, we find that the parameter values that maximize (16.15) are $\tilde{\beta}_1 = -1.1525$ and $\tilde{\beta}_2 = 0.1892$. These values maximize the log-likelihood function, $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$, and also maximize the likelihood function $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$. They are the *maximum likelihood estimates*. Any other values of the parameters that we might try will yield a lower value of the log-likelihood function. Plugging these values into (16.15), we obtain the value of the log-likelihood function evaluated at the maximum likelihood estimates, which is $L(\tilde{\beta}_1, \tilde{\beta}_2|\mathbf{y}, \mathbf{x}) = -1.5940$.

An interesting feature of the maximum likelihood estimation procedure is that while its properties in small samples are not known, we can show that in large samples the maximum likelihood estimator is normally distributed, consistent and *best*, in the sense that no competing estimator has smaller variance. The properties of maximum likelihood estimators are fully discussed in Appendix C.8.

We have used only three observations in the numerical illustration above for demonstration purposes only. In practice, such maximum likelihood estimation procedures should only be used when large samples are available. In the following section, we present another simple example that will demonstrate more aspects of the probit choice model.

EXAMPLE 16.4 | The Transportation Data: Probit

In Example 16.2, we estimated a linear probability model using the transportation data, *transport*. In this example, we carry out probit estimation. The probit model is $P(AUTO = 1) = \Phi(\beta_1 + \beta_2 DTIME)$. The maximum likelihood estimates of the parameters are

$$\begin{aligned}\tilde{\beta}_1 + \tilde{\beta}_2 DTIME &= -0.0644 + 0.3000 DTIME \\ \text{(se)} \qquad & \quad (0.3992) \quad (0.1029)\end{aligned}$$

The values in parentheses below the parameter estimates are estimated standard errors that are valid in large samples. These standard errors can be used to carry out hypothesis tests and construct interval estimates in the usual way, with the qualification that they are valid in large samples. The negative sign of $\tilde{\beta}_1$ implies that when commuting times via bus and auto are equal so $DTIME = 0$, individuals have a bias against driving to work, relative to public transportation, The estimated probability of a person choosing to drive to work when $DTIME = 0$ is $\hat{P}(AUTO = 1|DTIME = 0) = \Phi(-0.0644) = 0.4743$. The positive sign of $\tilde{\beta}_2$ indicates that an increase in public transportation travel time, relative to auto travel time, increases the probability that an individual will choose to drive to work, and this coefficient is statistically significant.

Suppose that we wish to estimate the marginal effect of increasing public transportation time, given that travel via public transportation currently takes 20 minutes longer than auto travel. Using (16.11),

$$\begin{aligned}\widehat{\frac{dp}{dDTIME}} &= \phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME)\tilde{\beta}_2 \\ &= \phi(-0.0644 + 0.3000 \times 2)(0.3000) \\ &= \phi(0.5355)(0.3000) = 0.3456 \times 0.3000 = 0.1037\end{aligned}$$

For the probit probability model, an incremental (10-minute) increase in the travel time via public transportation increases the probability of travel via auto by approximately 0.1037, given that taking the bus already requires 20 minutes more travel time than driving.

The estimated parameters of the probit model can also be used to "predict" the behavior of an individual who must choose between auto and public transportation to travel to work. If an individual is faced with the situation that it takes 30 minutes longer to take public transportation than to drive to work, then the estimated probability that auto transportation will be selected is calculated using (16.12):

$$\hat{p} = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME) = \Phi(-0.0644 + 0.3000 \times 3) = 0.7983$$

Since the estimated probability that the individual will choose to drive to work is 0.7983, which is greater than 0.5, we "predict" that when public transportation takes 30 minutes longer than driving to work, the individual will choose to drive.

EXAMPLE 16.5 | The Transportation Data: More Postestimation Analysis

In Example 16.4, we estimated the probit model for transportation choice and illustrated basic calculations. In this example, we carry out further, more advanced, postestimation analysis.

Marginal Effect at a Representative Value (MER)

The marginal effect of a change in the travel time differential is

$$\widehat{\frac{dp}{dDTIME}} = \phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME)\tilde{\beta}_2 = g(\tilde{\beta}_1, \tilde{\beta}_2)$$

The marginal effect is an estimator, since, given *DTIME*, it is a function of the estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The discussions of the "delta method" in Section 5.7.4 and Appendix 5B are relevant because the marginal effect is a *nonlinear* function of $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The marginal effect estimator is consistent and asymptotically normal with a variance given by equation (5B.4). Using this result, we can test marginal effects or compute interval estimates for them. For example, if the time differential is currently 20 minutes, so that the representative value is $DTIME = 2$, the estimated marginal effect (MER) is 0.1037 and the estimated standard error of the marginal effect is 0.0326 using the delta method. Therefore, a 95% interval estimate of the marginal effect, using the *t*-critical value $t_{(0.975,19)} = 2.093$, is [0.0354, 0.1720]. This interval is fairly wide. Recall, however, that the maximum likelihood estimates are based on only 21 observations, which is a very small sample. The details of the calculation of the standard error are given in Appendix 16A.1.

Marginal Effect at the Mean (MEM)

If particular values of interest are difficult to identify, many researchers evaluate the marginal effect "at the means," MEM. In these data, the average time travel differential is $\overline{DTIME} = -0.1224$ (1.2 minutes), and for this value, the marginal effect of a 10-minute increase in the time travel differential is 0.1191. The slightly larger effect, compared to $DTIME = 2$, is consistent with the second point in the Section 16.2.1 discussion. When the mean difference in travel time is near zero, the effect of a change in travel time difference is greater. We can compute a standard error for this marginal effect just as we did for MER, if we treat $\overline{DTIME}$ as given.

Average Marginal Effect (AME)

Rather than evaluate the marginal effect at a specific value, or the mean value, we can compute the average of the marginal effects evaluated at each sample data point. That is,

$$\widehat{\text{AME}} = \frac{1}{N}\sum_{i=1}^{N}\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)\tilde{\beta}_2 = \frac{1}{N}\tilde{\beta}_2\sum_{i=1}^{N}\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)$$

The average marginal effect has become a popular alternative to computing the marginal effect at the mean as it summarizes the response of individuals in the sample to a change in the value of an explanatory variable. For the current example, $\widehat{\text{AME}} = 0.0484$, which is the sample average estimated increase in probability given a 10-minute increase in bus travel time relative to auto travel time. Because the estimated marginal effect is different for each individual in the sample, we are interested in not only its average value but also its variation in the sample. The sample standard deviation of $\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)\tilde{\beta}_2$ is 0.0365, and its minimum and maximum values are 0.0025 and 0.1153.

We can evaluate the standard error of the average marginal effect using the delta method. Recall that $\widehat{\text{AME}} = 0.0484$. Its standard error estimated using the delta method is 0.0034. Details of this calculation are given in Appendix 16A.2. A 95% interval estimate of the population average marginal effect, using the *t*-critical value, is [0.0413, 0.0556]. This is much narrower than the MER interval estimate because we are estimating a different quantity, namely $\text{AME} = \frac{1}{N}\beta_2\sum_{i=1}^{N}\phi(\beta_1 + \beta_2 DTIME_i)$.

Estimated Probability of Driving

The estimated probability that $AUTO = 1$ given that the commuting time difference is 30 minute is calculated as $\hat{p} = \Phi\left(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME\right) = \Phi(-0.0644 + 0.3000 \times 3) = 0.7983$. Note that the predicted probability is a nonlinear function of the parameter estimates. Using the delta method, we can compute a standard error for the prediction and thus an interval estimate. The details of the calculation of the standard error are given in Appendix 16A.3. The calculated standard error is 0.1425, so that a 95% prediction interval, again using the t-critical value $t_{(0.975,19)} = 2.093$, is [0.5000, 1.0966]. Note that the upper endpoint of the interval is greater than 1, which means that some of the values are infeasible.

This example has been used to illustrate in a simple problem how probit works. In reality, estimating complicated models like probit and logit with as few observations as we are using, $N = 21$, is not a good idea. In fact, microeconometric models can have many more parameters and sometimes are estimated using very large data sets.

16.2.4 The Logit Model for Binary Choices

A frequently used alternative to the probit model for binary choice situations is the **logit** model. These models differ only in the particular S-shaped curve used to constrain probabilities to the [0, 1] interval. If L is a **logistic random variable**, then its *pdf* is

$$\lambda(l) = \frac{e^{-l}}{\left(1 + e^{-l}\right)^2}, \quad -\infty < l < \infty \tag{16.16}$$

The corresponding cumulative distribution function, unlike the normal distribution, has a closed-form expression, which makes analysis somewhat easier. The cumulative distribution function for a logistic random variable is

$$\Lambda(l) = p[L \le l] = \frac{1}{1 + e^{-l}} \tag{16.17}$$

In the logit model, if there is a single explanatory variable x, the probability $p(x)$ that the observed value y takes the value 1 is

$$p(x) = P\left[L \le \gamma_1 + \gamma_2 x\right] = \Lambda\left(\gamma_1 + \gamma_2 x\right) = \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} \tag{16.18}$$

A more generally useful form of $p(x)$ is

$$p(x) = \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} = \frac{\exp\left(\gamma_1 + \gamma_2 x\right)}{1 + \exp\left(\gamma_1 + \gamma_2 x\right)}$$

Then the probability that $y = 0$ is

$$1 - p(x) = \frac{1}{1 + \exp\left(\gamma_1 + \gamma_2 x\right)}$$

Represented in this way, the logit model can be extended to cases in which the choice is between more than two alternatives, as we will see in Section 16.3.

In maximum likelihood estimation of the logit model, the probability given in (16.18) is used to form the likelihood function (16.14) by inserting "Λ" for "Φ." To interpret the logit estimates, the equations (16.11) and (16.12) are still valid, using (16.16) instead of the normal *pdf*.

The shapes of the logistic and normal *pdfs* are somewhat different and maximum likelihood estimates of β_1 and β_2 will differ from γ_1 and γ_2. Roughly[4]

$$\tilde{\gamma}_{\text{Logit}} \cong 4\hat{\beta}_{\text{LPM}}$$
$$\tilde{\beta}_{\text{Probit}} \cong 2.5\hat{\beta}_{\text{LPM}}$$
$$\tilde{\gamma}_{\text{Logit}} \cong 1.6\tilde{\beta}_{\text{Probit}}$$

While the probit and logit parameter estimates differ, the marginal effects and predicted probabilities differ very little in most cases. In these expressions LPM denotes the linear probability model.

EXAMPLE 16.6 | An Empirical Example from Marketing

In Section 7.4.1, we introduced the example of a linear probability model for the choice between Coke and Pepsi. Here, we compare the linear probability model to the probit and logit models for this binary choice. The outcome variable is *COKE*

$$COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$$

The expected value of this variable is $E(COKE|\mathbf{x}) = p_{COKE}$ = probability that Coke is chosen. As explanatory variables, $\mathbf{x}$, we use the relative price of Coke to Pepsi (*PRATIO*), as well as *DISP_COKE* and *DISP_PEPSI*, which are indicator variables taking the value 1 if the respective store display is present and 0 if it is not present. We anticipate that the presence of a Coke display will increase the probability of a Coke purchase, and the presence of a Pepsi display will decrease the probability of a Coke purchase.

The data file *coke* contains "scanner" data on 1140 individuals who purchased Coke or Pepsi. The linear probability, probit, and logit models for the choice are

$$\begin{aligned} p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \alpha_1 + \alpha_2 PRATIO + \alpha_3 DISP\_COKE \\ &\quad + \alpha_4 DISP\_PEPSI \\ p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE \\ &\quad + \beta_4 DISP\_PEPSI) \\ p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \Lambda(\gamma_1 + \gamma_2 PRATIO + \gamma_3 DISP\_COKE \\ &\quad + \gamma_4 DISP\_PEPSI) \end{aligned}$$

We have given the choice model parameters different symbols to emphasize that the parameters have different meanings. The estimates are given in Table 16.1.

The parameters and their estimates vary across the models and no direct comparison is very useful. More relevant, however, is the comparison of the estimated probabilities and marginal effects implied by the alternative models.

TABLE 16.1 Coke-Pepsi Choice Models

	LPM	Probit	Logit
C	0.8902	1.1081	1.9230
	(0.0653)	(0.1900)	(0.3258)
PRATIO	−0.4009	−1.1460	−1.9957
	(0.0604)	(0.1809)	(0.3146)
DISP_COKE	0.0772	0.2172	0.3516
	(0.0339)	(0.0966)	(0.1585)
DISP_PEPSI	−0.1657	−0.4473	−0.7310
	(0.0344)	(0.1014)	(0.1678)

Standard errors in parentheses (White robust se for LPM)

Estimated probabilities at representative values Suppose that *PRATIO* = 1.1, indicating that the price of Coke is 10% higher than the price of Pepsi, and no store displays are present. Using the linear probability model, the estimated probability of Coke choice is 0.4493 with standard error 0.0202. Using probit, the estimated probability is 0.4394 with standard error 0.0218, and for logit, the estimated probability is 0.4323 with standard error 0.0224.

Average marginal effects (AME) In the linear probability model, the estimated marginal effect of *PRATIO* is −0.4009. This does not depend on the values of the variables. For the probit model, the average marginal effect of *PRATIO* is −0.4097 with standard error 0.0616, and for the logit model, the average marginal effect of *PRATIO* is −0.4333 with standard error 0.0639. In this example, the average marginal effect from the probit model is not too different from that implied by the linear probability model.

[4]T. Amemiya (1981) "Qualitative response models: A Survey," *Journal of Economic Literature*, 19, pp. 1483–1536, or A. Colin Cameron and Pravin K. Trivedi (2010) *Microeconometrics Using Stata: Revised Edition*, Stata Press, p. 465.

Marginal effect at a representative value (MER) If we examine specific scenarios then differences appear. For example, suppose *PRATIO* = 1.1, indicating that the price of Coke is 10% higher than the price of Pepsi, and no store displays are present. The estimated marginal effect of *PRATIO* from the probit model is −0.4519, with standard error 0.0703. Using the logit estimates, the marginal effect is −0.4898 with standard error 0.0753.

Prediction success Another basis for comparison is how well the alternative models predict choice outcomes. For the linear probability model, compute the predicted value $\widehat{COKE}$, then predict consumer choice by comparing this value to 0.5. If $\widehat{COKE}$ is greater than 0.5, we predict the consumer will choose Coke. For the probit model, we estimate the probability of choosing Coke using equation (16.10). Using the 0.5 threshold, we find that of the 510 consumers who chose *COKE*, 247 were correctly predicted. Of the 630 who chose *PEPSI*, 507 were correctly predicted. In this example, the number of correct predictions is identical for the linear probability model, probit and logit.

16.2.5 Wald Hypothesis Tests

Hypothesis tests concerning individual coefficients in probit and logit models are carried out in the usual way based on an "asymptotic-*t*" test. If the null hypothesis is $H_0: \beta_k = c$, then the test statistic using the probit model is

$$t = \frac{\tilde{\beta}_k - c}{\text{se}(\tilde{\beta}_k)} \overset{a}{\sim} N(0,1)$$

where $\tilde{\beta}_k$ is the probit parameter estimator. The test is asymptotically justified and we should use the test critical values from the standard normal distribution. For two-tail tests, these are the familiar 1.645 for 10%, 1.96 for 5%, and 2.58 for 1%. However, it is not uncommon to take a more conservative approach and, if the sample size is not very large, to use critical values from the $t_{(N-K)}$ distribution, where K is the number of parameters estimated. Your software may report "z" statistics instead of "t" and automatically compute p-values and calculate interval estimates with critical numbers from the standard normal distribution, rather than the t-distribution.

The t-test is based on the *Wald principle*, which uses the model coefficient estimates, estimated variances, covariances, and standard errors that are asymptotically valid. This testing principle is discussed in Appendix C.8.4. It is common for software packages to have "built in" Wald test statements (something like "TEST") that are convenient to use after a model is estimated. For linear hypotheses, such as $H_0: c_2\beta_2 + c_3\beta_3 = c_0$, the test statistic is of the familiar form,

$$t = \frac{\left(c_2\tilde{\beta}_2 + c_3\tilde{\beta}_3\right) - c_0}{\sqrt{c_2^2\widehat{\text{var}}\left(\tilde{\beta}_2\right) + c_3^2\widehat{\text{var}}\left(\tilde{\beta}_3\right) + 2c_2c_3\widehat{\text{cov}}\left(\tilde{\beta}_2, \tilde{\beta}_3\right)}}$$

If the null hypothesis is true, then this statistic has an asymptotic $N(0, 1)$ distribution but again $t_{(N-K)}$ might be used if the sample is not truly large. For **joint linear hypotheses**, such as

$$H_0: c_2\beta_2 + c_3\beta_3 = c_0, \quad a_4\beta_4 + a_5\beta_5 = a_0$$

a valid large sample Wald test is based on the chi-square distribution. If there are J joint hypotheses, the Wald statistic has an asymptotic $\chi^2_{(J)}$ distribution. The null hypothesis is rejected if the Wald test statistic, W, is greater than or equal to the $(1 - \alpha)$ percentile of the $\chi^2_{(J)}$ distribution, $\chi^2_{(1-\alpha,J)}$. In Section 6.1.5, we discuss large sample tests in the linear regression model. The chi-square test was labeled $\hat{V}_1$ in equation (6.14), and it was calculated as the difference between the sums of squared residuals from an unrestricted and a restricted model, divided by the estimated error variance. That is *not* the way the statistic is calculated in nonlinear models such as probit and logit, but the interpretation is the same. There is a "small-sample" conservative correction using the F-statistic, $F = W/J \overset{a}{\sim} F_{(J,N-K)}$, which is similar to using t-critical values instead of those from the $N(0, 1)$ distribution. Do not be surprised if your software reports a chi-square statistic instead of a t-statistic even when only one hypothesis is being tested.

EXAMPLE 16.7 | Coke Choice Model: Wald Hypothesis Tests

Here are some examples of various tests in the Coke choice model.

Test of significance Using the estimates in Table 16.1, we can test the significance of the coefficients in the usual way. The probit model for *COKE* is

$$p_{COKE} = \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI)$$

We might like to test the null hypothesis $H_0: \beta_3 \leq 0$ against $H_1: \beta_3 > 0$. The test statistic is $t = \tilde{\beta}_3/\text{se}(\tilde{\beta}_3) \overset{a}{\sim} N(0,1)$ if the null hypothesis is true. Using a 5% one-tail test, the critical value is $z_{(0.95)} = 1.645$. The calculated value of the test statistic is $t = \tilde{\beta}_3/\text{se}(\tilde{\beta}_3) = 2.2481$, and thus, we reject the null hypothesis at the 5% level and conclude that a display for Coke has a positive effect on the probability that a consumer will purchase Coke. Using a TEST statement might also produce the Wald statistic $W = 5.0540$. For a single hypothesis $W = t^2$. The Wald test statistic is designed for two-tail tests; in this case $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$, yields a two-tail p-value of $p = 0.0246$. If your software reports a t-statistic or an F-statistic, the p-value will be slightly larger, $p = 0.0248$. There is little difference here because the sample is large with $N = 1140$ observations. The Wald test critical value is $\chi^2_{(0.95,1)} = 3.841$ from Statistical Table 3.

Testing an economic hypothesis Another hypothesis of interest is $H_0: \beta_3 = -\beta_4$ versus $H_1: \beta_3 \neq -\beta_4$. This hypothesis is that the coefficients on the display variables are equal in magnitude but opposite in sign or that the effects of the Coke and Pepsi displays have an equal but opposite effect on the probability of choosing Coke. The t-test statistic is

$$t = \frac{\tilde{\beta}_3 + \tilde{\beta}_4}{\text{se}(\tilde{\beta}_3 + \tilde{\beta}_4)} \overset{a}{\sim} N(0,1)$$

Noting that it is a two-tail alternative hypothesis, we reject the null hypothesis at the $\alpha = 0.05$ level if $t \geq 1.96$ or $t \leq -1.96$. The calculated t-value is $t = -2.3247$, so we reject the null hypothesis and conclude that the effects of the Coke and Pepsi displays are not of equal magnitude with opposite sign. This test is asymptotically valid because $N - K = 1140 - 4 = 1136$ is a large sample. Automatic TEST statements usually generate the chi-square distribution version of the test, which in this case is the square of the t-statistic, $W = 5.4040$. The 5% critical value is $\chi^2_{(0.95,1)} = 3.841$ so we reject the null hypothesis. We reach the same conclusion as using the t-test. The link between the t- and chi-square test is fully explained in Appendix C.8.4.

Testing joint significance Another hypothesis of interest is

$$H_0: \beta_3 = 0,\ \beta_4 = 0 \qquad H_1: \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

This joint null hypothesis is that neither the Coke nor Pepsi display affects the probability of choosing Coke. Here we are testing $J = 2$ hypotheses, so that the Wald statistic has an asymptotic $\chi^2_{(2)}$ distribution. Using Statistical Table 3, the 0.95 percentile value for this distribution is 5.991. In this case, the value of the Wald statistic is $W = 19.4594$, and thus, we reject the null hypothesis and conclude that the Coke or Pepsi display has an effect on the probability of choosing Coke. This test statistic value can be computed using the automatic TEST statement in your software.

Testing the overall model significance As in the linear regression model, we are interested in testing the overall significance of the probit model. In the Coke choice example, the null hypothesis for this test is $H_0: \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is that at least one of the parameters is not zero. The value of the Wald test statistic is 132.54. The test statistic has an asymptotic $\chi^2_{(3)}$ distribution if the null hypothesis is true. The 0.95 percentile value for this distribution is 7.815, so we reject the null hypothesis that none of the explanatory variables help explain the choice of Coke versus Pepsi.

16.2.6 Likelihood Ratio Hypothesis Tests

When using maximum likelihood estimators, such as probit and logit, tests based on the **likelihood ratio principle** are generally preferred. Appendix C.8.4 contains a discussion of this methodology. The idea is much like the F-test in the linear regression model. One test component is the log-likelihood function value in the unrestricted, full model (call it $\ln L_U$) evaluated at the maximum likelihood estimates. This calculation was illustrated in Example 16.3. Whenever a model is estimated by maximum likelihood, the maximized value of the log-likelihood function is automatically reported by econometric software. The second ingredient in a likelihood ratio test is the log-likelihood function value from the model that is "restricted" by imposing the condition that the null hypothesis is true (call it $\ln L_R$). Thus, the likelihood ratio test has the disadvantage of requiring two estimations of the model; once for the original model and once for the model that assumes the hypothesis is true. The likelihood ratio test statistic is $LR = 2(\ln L_U - \ln L_R)$.

The idea is that if the null hypothesis is true, then there should be little difference between the log-likelihood function with or without the hypothesis being assumed true. In that case, the *LR* statistic will be small but always greater than zero. If the null hypothesis is not true, then when we estimate the model assuming that it is true, the model should not fit as well, and the maximum value of the restricted log-likelihood function will be lower, making *LR* larger. Large values of the *LR* test statistic are evidence against the null hypothesis. If the null hypothesis is true, the statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of hypotheses, J, being tested. The null hypothesis is rejected if the value *LR* is larger than the chi-square distribution critical value, $\chi^2_{(1-\alpha,J)}$.

EXAMPLE 16.8 | Coke Choice Model: Likelihood Ratio Hypothesis Tests

We can use likelihood ratio tests for the same hypotheses considered in Example 16.7.

Test of significance The probit model for *COKE* is

$$p_{COKE} = \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI)$$

To test the null hypothesis $H_0: \beta_3 = 0$ against $H_1: \beta_3 \neq 0$ using the likelihood ratio principle, we first note that the maximized value of the log-likelihood function is $\ln L_U = -710.9486$. If the null hypothesis is true, then the restricted model is $p_{COKE} = \Phi(\beta_1 + \beta_2 PRATIO + \beta_4 DISP\_PEPSI)$. Estimating this model by maximum likelihood, we find $\ln L_R = -713.4803$, which is smaller than in the original model, as it must be. Imposing constraints on a probit model will reduce the maximized value of the log-likelihood function. Then

$$LR = 2(\ln L_U - \ln L_R) = 2[-710.9486 - (-713.4803)] = 5.0634$$

The 5% critical value is $\chi^2_{(0.95,1)} = 3.841$. We reject the null hypothesis that a display for Coke has no effect.

Test of an economic hypothesis To test $H_0: \beta_3 = -\beta_4$, we first obtain the unrestricted probit model log-likelihood value, $\ln L_U = -710.9486$. The restricted probit model is obtained by imposing the condition $\beta_3 = -\beta_4$ on the model, leading to

$$\begin{aligned} p_{COKE} &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI) \\ &= \Phi(\beta_1 + \beta_2 PRATIO - \beta_4 DISP\_COKE + \beta_4 DISP\_PEPSI) \\ &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_4 (DISP\_PEPSI - DISP\_COKE)) \end{aligned}$$

Estimating this model by maximum likelihood probit, we obtain $\ln L_R = -713.6595$. The likelihood ratio test statistic value is then

$$LR = 2(\ln L_U - \ln L_R) = 2[-710.9486 - (-713.6595)] = 5.4218$$

This value is larger than the 0.95 percentile from the $\chi^2_{(1)}$ distribution, $\chi^2_{(0.95,1)} = 3.841$. Note that the values of the *LR* and Wald statistics (from Example 16.7) are not the same but are close in this case. The Wald test statistic value is easier to compute, since it requires only the maximum likelihood estimates for the original, unrestricted model. However, the likelihood ratio test has been found to be more reliable in a wide variety of more complex testing situations, and it is the preferred test.[5]

Test of joint significance To test the joint null hypothesis $H_0: \beta_3 = 0,\ \beta_4 = 0$, use the restricted model $E(COKE|\mathbf{x}) = \Phi(\beta_1 + \beta_2 PRATIO)$. The value of the likelihood ratio test statistic is 19.55, which is larger than the $\chi^2_{(2)}$ 0.95 percentile value 5.991. We reject the null hypothesis that neither the Coke nor Pepsi display has an effect on the choice of Coke.

Testing the overall model significance As in the linear regression model, we are interested in testing the overall significance of the probit model. In the Coke choice example, the null hypothesis for this test is $H_0: \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is that at least one of the parameters is not zero. If the null hypothesis is true, the restricted model is $E(COKE) = \Phi(\beta_1)$. The log-likelihood value for this restricted model is $\ln L_R = -783.8603$ and the value of the likelihood ratio test statistic is $LR = 145.8234$. The test statistic has an asymptotic $\chi^2_{(3)}$ distribution if the null hypothesis is true. The 0.95 percentile value for this distribution is 7.815, so we reject the null hypothesis that none of the explanatory variables help explain the choice of Coke versus Pepsi. In addition, like in the linear regression model, this "overall" test is reported in standard probit computer output.

[5] Griffiths, W. E., Hill, R. C., & Pope, P. (1987). Small Sample Properties of Probit Model Estimators. *Journal of the American Statistical Association*, 82, 929–937.

16.2.7 Robust Inference in Probit and Logit Models

You may be wondering if there are "robust" standard errors for use with probit and logit that correct for heteroskedasticity and/or serial correlation. Unfortunately, the answer is no. As noted in Chapter 8, equation (8.32), the 0-1 random variable y_i has conditional variance $\text{var}(y_i|\mathbf{x}_i) = p(\mathbf{x}_i)[1 - p(\mathbf{x}_i)]$. In the probit model, for example, this means that

$$\text{var}(y_i|\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\left[1 - \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\right]$$

There is no other possible variance if the probit model is correct. Maximum likelihood estimation of the probit model does not require any adjustment for this built in heteroskedasticity. Some software packages do have probit with a robust option, but it does not provide the type of robust results we have seen in Chapters 8 and 9. If you happen to use one of these options, and if the "robust" standard errors are much different from the usual probit standard errors, then, if anything, it is a symptom of some specification problem, such as incorrect functional form.

An exception is when there are data clusters. In Section 15.2.1, we introduced cluster-robust standard errors. There we discussed clusters in the context of panel data. However, clusters of observations, in which there are intracluster correlations, can occur in many contexts. We may observe individuals within different villages, and there may be a common unobserved heterogeneity within villages representing a "village effect." The unobserved heterogeneity causes a correlation among individuals in the same village, while there is no correlation among individuals across villages. In these situations, conventional standard errors may greatly overstate the precision of estimation. Therefore, using cluster-robust standard errors with probit and logit is recommended when the problem is suitable. In general, this means that there are many clusters with not too many observations in each.[6] Be careful when implementing cluster-robust standard errors as the computer command may be quite different from the usual "robust" standard error command.

16.2.8 Binary Choice Models with a Continuous Endogenous Variable

There are several ways that probit concepts can be combined with endogenous variables. The first is when the outcome variable is binary, as in the linear probability or probit models, and an explanatory variable is endogenous. As in our discussions of instrumental variables and two-stage least squares estimation in Chapters 10 and 11, the estimation methods here require instrumental variables.

The first, and easiest, option is to estimate a linear probability model for the binary outcome variable using IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where $y_{i1} = 1$ or 0, y_{i2} is a continuous endogenous variable, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

Using the IV/2SLS estimation approach, we first estimate this equation by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimating this model by OLS produces IV/2SLS estimates. However, as always, to obtain correct standard errors use IV/2SLS software, and in this case use heteroskedasticity robust standard errors.

[6] A complete but advanced resource is A. Colin Cameron and Douglas L. Miller (2015). A Practitioner's Guide to Cluster-Robust Inference, *The Journal of Human Resources*, 50(2), 317–372.

This approach is familiar and easy to implement. As always we must be concerned about the strength of the instrumental variable. The coefficient π_3 must not be zero, and when the first-stage model is estimated, it must be statistically very significant. As previously noted, using the linear probability model is not ideal when the outcome variable is binary. The procedure we have outlined ignores the binary character of the outcome variable, but it may reasonably estimate the population average marginal effect. There is another, more theoretically complicated, maximum likelihood estimator that is called *instrumental variables probit*, or simply *IV probit*.[7] This estimator is available in some software packages.

EXAMPLE 16.9 | Estimating the Effect of Education on Labor Force Participation

When studying the wages of married women, Examples 10.1–10.7 using data file *mroz*, we were very concerned with the endogeneity of education. In those examples, we only considered women who were in the labor force and had an observable market wage. Now we ask about the effect of education on the decision to join the labor force or not. Let

$$LFP = \begin{cases} 1 & \text{in labor force} \\ 0 & \text{not in labor force} \end{cases}$$

Consider the linear probability model

$$LFP = \alpha_1 EDUC + \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 KIDSL6 + \beta_5 AGE + e$$

Suppose the instrumental variable for *EDUC* is *MOTHEREDUC*. The first-stage equation is

$$EDUC = \pi_1 + \pi_2 EXPER + \pi_3 EXPER^2 + \pi_4 KIDSL6 + \pi_5 AGE + \pi_6 MOTHEREDUC + v$$

In the first-stage estimation, the t-value for the coefficient of *MOTHEREDUC* is 12.85, which using conventional standards indicates that this instrument is not weak. The two-stage least squares estimates of the labor force participation equation, with robust standard errors, are

$$\widehat{LFP} = 0.0388 EDUC + 0.5919 + 0.0394 EXPER$$
$$\text{(se)} \quad (0.0165) \qquad (0.2382) \quad (0.0060)$$
$$- 0.0006 EXPER^2 - 0.2712 KIDSL6 - 0.0177 AGE$$
$$(0.0002) \qquad (0.03212) \qquad (0.0023)$$

We estimate that each additional year of education increases the probability of a married woman being in the labor force by 0.0388, holding all else constant. The regression-based Hausman test for the endogeneity of education, using robust standard errors, has a p-value of 0.646. Thus, we cannot reject the exogeneity of education in this model, using the instrument *MOTHEREDUC*.

16.2.9 Binary Choice Models with a Binary Endogenous Variable

Modify the model in Section 16.2.8 so that the endogenous variable y_{i2} is binary. The first, and easiest, option is again to estimate a linear probability model for the binary outcome variable using IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where $y_{i1} = 1$ or 0, $y_{i2} = 1$ or 0, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

Using the IV/2SLS estimation approach, we first estimate this equation by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we

[7] See William Greene (2018) *Econometric Analysis, Eighth Edition*, Prentice-Hall, page 773, or Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, p. 585–594. These references are very advanced.

have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimate this model by OLS to obtain IV/2SLS estimates. Of course, as always, use proper IV/2SLS software and, because the dependent variable is binary, use heteroskedasticity robust standard errors.

It is tempting but *incorrect* to think that the first-stage equation can be estimated by probit, followed by substituting $\tilde{p}_i = \tilde{P}(y_{i2} = 1) = \Phi(\tilde{\pi}_1 + \tilde{\pi}_2 x_{i2} + \tilde{\pi}_3 z_i)$ into the equation of interest, and then applying either probit or the linear probability model. The second estimation is called a forbidden regression.[8] Two-stage least squares works only when it consists of two OLS regressions, substituting OLS fitted values from a first-stage regression in for the endogenous variable in the first equation. 2SLS works because OLS has the property that the residuals are uncorrelated with the explanatory variables.

Once again the linear probability model approach "works" but does not use the fact that $y_{i1} = 1$ or 0 and $y_{i2} = 1$ or 0 are binary variables. A maximum likelihood estimation approach called bivariate probit[9] does take this into account.

EXAMPLE 16.10 | Women's Labor Force Participation and Having More Than Two Children

The Angrist and Evans (1998)[10] model of labor force participation, $LFP = 1$ or 0, includes as an explanatory variable the indicator variable $MOREKIDS = 1$ if the woman has three or more children, and $MOREKIDS = 0$ otherwise. Intuitively, we think having three or more children will have a negative effect on the probability of labor force participation. The very clever instrumental variable used is the indicator variable where the value $SAMESEX = 1$ if the woman's first two children are of the same sex, and $SAMESEX = 0$ otherwise. The idea behind this instrumental variable is that while it should have no direct effect on labor force participation it is correlated with a woman having three or more children. If a woman's first two children are both boys (girls), then she may be inclined to have another child in the hope of getting a girl (boy).

16.2.10 Binary Endogenous Explanatory Variables

Modify the model in Section 16.2.9 so that the outcome variable y_{i1} is continuous and the endogenous variable y_{i2} is binary. This model has long been studied and was first called a *dummy endogenous variable model* by Nobel prize winner James Heckman. The first, and easiest, option is to use IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where y_{i1} is continuous, the endogenous variable $y_{i2} = 1$ or 0, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

[8] Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, p. 267–268 and 596–597.

[9] See William Greene (2018) *Econometric Analysis, Eighth Edition*, Prentice-Hall, Chapter 17.9, or Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data*, Second Edition, MIT Press, pages. 594–599. These references are very advanced.

[10] Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size, *The American Economic Review*, Vol. 88, No. 3 (Jun., 1998), pp. 450–477. See also Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, pp. 597–598.

Using the IV/2SLS estimation approach, we first estimate this linear probability model by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimating this model by OLS yields the 2SLS estimates. As always proper IV/2SLS software should be used.

The presence of an endogenous binary variable is an important feature in some **treatment effect** models.[11]

EXAMPLE 16.11 | Effect of War Veteran Status on Wages

A widely cited work by Joshua Angrist examines the effect of serving in the Vietnam war on the wages of male American workers. December 1, 1969, there was a lottery to determine eligibility for being drafted into service. Imagine 366 slips of paper each written with a birth date. The slips are placed in a jar, mixed up, and a slip drawn. The first date drawn was September 14. All men of eligible age with that birthday were given draft lottery number 1. The second date drawn was April 24 and was given lottery number 2, and so on. In the first lottery, all those with lottery numbers 195 or less were called to report for possible induction into the military. Some of those chosen did not serve for medical or other reasons, and some chose to volunteer. Thus, those who ultimately served, and became war veterans, did not correspond exactly to those with lottery numbers less than or equal to 195.

Consider a model of worker earnings, 10 years after the draft. Let *VETERAN* = 1 if a person was a veteran and = 0 otherwise. Because some chose to volunteer, the binary variable *VETERAN* is endogenous in the model

$$EARNINGS = \alpha_2 VETERAN + \beta_1 + \beta_2 OTHER\_FACTORS + e_i$$

What is a possible instrument? A person's lottery number is correlated with veteran status. More specifically, let *LOTTERY* = 1 if a person's draft lottery number was 195 or less, and *LOTTERY* = 0 otherwise. We anticipate that *LOTTERY* will be positively correlated with *VETERAN* and is a potential instrument. This type of binary IV leads to the Wald estimator, introduced in Exercises 10.5 and 10.6. The results of the IV estimation show that serving in the military has a negative and significant effect on wages.

16.2.11 Binary Choice Models and Panel Data

In Chapter 15, we used panel data to control for unobservable heterogeneity across individuals. The *fixed effects estimator* includes an indicator, or dummy, variable for each individual. Equivalently, the *within* estimator uses deviations about individual means to estimate coefficients of the regression function. We use the fixed effects estimator when the unobservable heterogeneity is correlated with the explanatory variables. The *random effects estimator* is a generalized least squares estimator that accounts for intra-individual error correlations caused by unobserved heterogeneity. It is more efficient than the fixed effects estimator but is inconsistent if the unobservable heterogeneity is correlated with any of the included explanatory variables.

If the outcome variable is binary, then using the panel data methods with the linear probability model is exactly the same as with the linear regression model. If there is unobserved heterogeneity that is correlated with one or more explanatory variables, then using the fixed effects estimator or the first difference estimator is appropriate. If the unobserved heterogeneity is not correlated with any explanatory variables, then using the random effects estimator is an option, as is the less efficient but consistent OLS estimator with robust cluster-corrected standard errors.

Using probit or logit with panel data is a different story. The probit model is a nonlinear model, that is, a nonlinear function of the parameters. If the unobserved heterogeneity is

[11] A discussion of the results and similar estimators can be found in Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Guide*, Princeton Press, pages 128–138. This reference is advanced. Other examples and estimation approaches for treatment effects are in Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21. This reference is very advanced. For an advanced and exhaustive survey see G. W. Imbens and J. M. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

correlated with the explanatory variables, we have a problem. The usual fixed effect approach to dealing with individual heterogeneity fails. If there are N individuals and $N \to \infty$ (gets large) while T remains fixed, then adding an indicator variable for each individual leads to a model in which the number of parameters we must estimate $N + K$ also approaches ∞. The probit estimator is no longer consistent because there are too many parameters. In statistics, this is called the *incidental parameters problem*. In the linear regression model, we avoid this problem by using the within-transformation, based on the Frisch–Waugh–Lovell theorem, so that we can estimate the regression function parameters without having to estimate all the fixed effects coefficients. This does not work in probit because of the nonlinear nature of the problem. In probit, we cannot apply the Frisch–Waugh–Lovell theorem and simply using variables in deviations about the mean form does not work. There is no fixed effects probit estimator, although researchers are considering methods for reducing the bias of the estimator so that it might be used. On the other hand, there is a type of panel logit fixed effects model called **conditional logit**, or sometimes *Chamberlain's conditional logit*,[12] recognizing the innovative econometrician Gary Chamberlain. It is not the same as introducing indicator variables for each individual into the logit model.

The probit model can however be combined with random effects to obtain a *random effects probit* model. The actual method of maximizing the likelihood function requires some tricky integrals, which can be solved using numerical approximations or simulations. As with the linear regression model, the random effects estimator is inconsistent if the random effects are correlated with the explanatory variables. It has been suggested that controls for time invariant factors, such as the time averages of the independent variables, $\bar{x}_{i\cdot}$, be introduced, similar to the Mundlak method for carrying out the Hausman test discussed in Chapter 15. The resulting model is called the *Mundlak–Chamberlain-correlated random effects probit model*.[13] The added variables $\bar{x}_{i\cdot}$ act like control variables, possibly reducing the random effects probit estimator bias.

A dynamic binary choice model, which includes the lagged value of the choice variable on the right-hand side as an explanatory variable, is an obvious way to handle habit persistence. Coke drinkers buying soda today are more likely to purchase Coke if they purchased Coke when shopping on the previous occasion. However, in such models, the lagged endogenous variable will be correlated with the random effect, as noted in Chapter 15. In this case, the previous estimators are inconsistent and new methods[14] must be considered.

16.3 Multinomial Logit

In probit and logit models, the decision-maker chooses between two alternatives. Clearly, we are often faced with choices involving more than two alternatives. These are called **multinomial choice** situations. Examples include the following:

- If you are shopping for a laundry detergent, which one do you choose? Tide, Cheer, Arm & Hammer, Wisk, and so on. The consumer is faced with a wide array of alternatives. Marketing researchers relate these choices to prices of the alternatives, advertising, and product characteristics.
- If you enroll in the business school, will you major in economics, marketing, management, finance, or accounting?
- If you are going to a mall on a shopping spree, which mall will you go to, and why?
- When you graduated from high school, you had to choose between not going to college and going to a private 4-year college, a public 4-year college, or a 2-year college. What factors led to your decision among these alternatives?

[12] See Wooldridge (2010, 620–622), Greene (2018, 787–789), or Baltagi (2013, 240–243). The material is advanced.

[13] See Greene (2018, 792–793) and Wooldridge (2010, 616–619).

[14] See Greene (2018, 794–796), Baltagi (2013, 248–253), and Wooldridge (2010, 625–630).

It would not take you long to come up with other illustrations. In each of these cases, we wish to relate the observed choice to a set of explanatory variables. More specifically, as in probit and logit models, we wish to explain and estimate the probability that an individual with a certain set of characteristics chooses one of the alternatives. The estimation and interpretation of such models is, in principle, similar to that in logit and probit models. The models themselves go under the names **multinomial logit**, **conditional logit**, and **multinomial probit**. We will discuss the most commonly used logit models.

16.3.1 Multinomial Logit Choice Probabilities

Suppose that a decision-maker must choose between several distinct alternatives. Let us focus on a problem with $J = 3$ alternatives. An example might be the choice facing a high-school graduate. Shall I attend a 2-year college, a 4-year college, or not go to college? The factors affecting this choice might include household income, the student's high-school grades, family size, race, and sex, and the parents' education. As in the logit and probit models, we will try to explain the probability that the *i*th person will choose alternative *j*,

$$p_{ij} = P[\text{individual } i \text{ chooses alternative } j]$$

In our example, there are $J = 3$ alternatives, denoted by $j = 1, 2$, or 3. These numerical values have no meaning because the alternatives in general have no particular ordering and are assigned arbitrarily. You can think of them as categories A, B, and C.

If we assume a single explanatory factor, x_i, then, in the multinomial logit specification, the probabilities of individual i choosing alternatives $j = 1, 2, 3$ are

$$p_{i1} = \frac{1}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 1 \tag{16.19a}$$

$$p_{i2} = \frac{\exp(\beta_{12} + \beta_{22}x_i)}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 2 \tag{16.19b}$$

$$p_{i3} = \frac{\exp(\beta_{13} + \beta_{23}x_i)}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 3 \tag{16.19c}$$

The parameters β_{12} and β_{22} are specific to the second alternative and β_{13} and β_{23} are specific to the third alternative. The parameters specific to the first alternative are set to zero to solve an identification problem and to make the probabilities sum to one.[15] Setting $\beta_{11} = \beta_{21} = 0$ leads to the 1 in the numerator of p_{i1} and the 1 in the denominator of each part of (16.19). Specifically, the term that would be there is $\exp(\beta_{11} + \beta_{21}x_i) = \exp(0 + 0x_i) = 1$.

A distinguishing feature of the multinomial logit model in (16.19) is that there is a single explanatory variable that describes the individual, *not* the alternatives facing the individual. Such variables are called individual specific. To distinguish the alternatives, we give them different parameter values. This situation is common in the social sciences, where surveys record many characteristics of the individuals, and choices they made.

16.3.2 Maximum Likelihood Estimation

Let y_{i1}, y_{i2}, and y_{i3} be indicator variables representing the choice made by individual i. If alternative 1 is selected, then $y_{i1} = 1$, $y_{i2} = 0$, and $y_{i3} = 0$. If alternative 2 is selected, then $y_{i1} = 0$, $y_{i2} = 1$, and $y_{i3} = 0$. In this model, each individual must choose one, and only one, of the available alternatives.

[15] Some software may choose the parameters of the last (Jth) alternative to set to zero, or perhaps the most frequently chosen group. Check your software documentation.

Estimation of this model is by maximum likelihood. Suppose that we observe three individuals, who choose alternatives 1, 2, and 3, respectively. Assuming that their choices are independent, then the probability of observing this outcome is

$$\begin{aligned} P(y_{11}=1, y_{22}=1, y_{33}=1 \mid x_1, x_2, x_3) &= p_{11} \times p_{22} \times p_{33} \\ &= \frac{1}{1+\exp(\beta_{12}+\beta_{22}x_1)+\exp(\beta_{13}+\beta_{23}x_1)} \\ &\quad \times \frac{\exp(\beta_{12}+\beta_{22}x_2)}{1+\exp(\beta_{12}+\beta_{22}x_2)+\exp(\beta_{13}+\beta_{23}x_2)} \\ &\quad \times \frac{\exp(\beta_{13}+\beta_{23}x_3)}{1+\exp(\beta_{12}+\beta_{22}x_3)+\exp(\beta_{13}+\beta_{23}x_3)} \\ &= L(\beta_{12}, \beta_{22}, \beta_{13}, \beta_{23}) \end{aligned}$$

In the last line, we recognize that this joint probability depends on the unknown parameters and is in fact the likelihood function. Maximum likelihood estimation seeks those values of the parameters that maximize the likelihood or, more specifically, the **log-likelihood function**, which is easier to work with mathematically. In a real application, the number of individuals will be greater than three, and computer software will be used to maximize the log-likelihood function numerically. While the task might look daunting, finding the maximum likelihood estimates in this type of model is fairly simple.

16.3.3 Multinomial Logit Postestimation Analysis

Given that we can obtain maximum likelihood estimates of the parameters, which we denote as $\tilde{\beta}_{12}$, $\tilde{\beta}_{22}$, $\tilde{\beta}_{13}$, and $\tilde{\beta}_{23}$, what can we do then? The first thing we might do is estimate the probability that an individual will choose alternative 1, 2, or 3. For the value of the explanatory variable x_0, we can calculate the predicted probabilities of each outcome being selected using (16.19). For example, the probability that such an individual will choose alternative 1 is

$$\tilde{p}_{01} = \frac{1}{1+\exp(\tilde{\beta}_{12}+\tilde{\beta}_{22}x_0)+\exp(\tilde{\beta}_{13}+\tilde{\beta}_{23}x_0)}$$

The estimated probabilities for alternatives 2 and 3, $\tilde{p}_{02}$ and $\tilde{p}_{03}$, can similarly be obtained. If we wanted to predict which alternative would be chosen, we might choose to predict that alternative j will be chosen if $\tilde{p}_{0j}$ is the maximum of the estimated probabilities.

Because the model is such a complicated nonlinear function of the parameters, it will not surprise you to learn that the βs are not "slopes." In these models, the **marginal effect** is the effect of a change in x, everything else held constant, on the probability that an individual chooses alternative $m = 1, 2$, or 3. It can be shown[16] that

$$\left.\frac{\Delta p_{im}}{\Delta x_i}\right|_{\text{all else constant}} = \frac{\partial p_{im}}{\partial x_i} = p_{im}\left[\beta_{2m} - \sum_{j=1}^{3}\beta_{2j}p_{ij}\right] \tag{16.20}$$

Recall that the model we are discussing has a single explanatory variable, x_i, and that $\beta_{21} = 0$.

[16] One can quickly become overwhelmed by the mathematics when seeking references on this topic. Two relatively friendly sources with good examples are *Regression Models for Categorical and Limited Dependent Variables* by J. Scott Long (Thousand Oaks, CA: Sage Publications, 1997) [see Chapter 5] and *Quantitative Models in Marketing Research* by Philip Hans Franses and Richard Paap (Cambridge University Press, 2001) [see Chapter 5]. At a much more advanced level, see *Econometric Analysis, Eighth edition* by William Greene (Upper Saddle River, NJ: Pearson Prentice Hall, 2018) [see Section 18.2.3].

Alternatively, and somewhat more simply, the difference in probabilities can be calculated for two specific values of x_i. If x_a and x_b are two values of x_i, then the estimated change in probability of choosing alternative 1 [$m = 1$] when changing from x_a to x_b is

$$\begin{aligned}\widetilde{\Delta p_1} &= \tilde{p}_{b1} - \tilde{p}_{a1} \\ &= \frac{1}{1 + \exp\left(\tilde{\beta}_{12} + \tilde{\beta}_{22} x_b\right) + \exp\left(\tilde{\beta}_{13} + \tilde{\beta}_{23} x_b\right)} \\ &\quad - \frac{1}{1 + \exp\left(\tilde{\beta}_{12} + \tilde{\beta}_{22} x_a\right) + \exp\left(\tilde{\beta}_{13} + \tilde{\beta}_{23} x_a\right)}\end{aligned}$$

This approach is good if there are certain scenarios that you as a researcher have in mind as typical or important cases or if x is an indicator variable with only two values, $x_a = 0$ and $x_b = 1$.

Another useful interpretive device is the **probability ratio**. It shows how many times more likely category j is to be chosen relative to the first category and is given by

$$\frac{P(y_i = j)}{P(y_i = 1)} = \frac{p_{ij}}{p_{i1}} = \exp\left(\beta_{1j} + \beta_{2j} x_i\right), \quad j = 2, 3 \tag{16.21}$$

The effect on the probability ratio of changing the value of x_i is given by the derivative

$$\frac{\partial\left(p_{ij}/p_{i1}\right)}{\partial x_i} = \beta_{2j} \exp\left(\beta_{1j} + \beta_{2j} x_i\right), \quad j = 2, 3 \tag{16.22}$$

The value of the exponential function $\exp\left(\beta_{1j} + \beta_{2j} x_i\right)$ is always positive. Thus, the sign of β_{2j} tells us whether a change in x_i will make the jth category more or less likely relative to the first category.

An interesting feature of the probability ratio (16.21) is that it does not depend on how many alternatives there are in total. There is the implicit assumption in logit models that the probability ratio between any pair of alternatives is **independent of irrelevant alternatives (IIA)**. This is a strong assumption, and if it is violated, multinomial logit may not be a good modeling choice. It is especially likely to fail if several alternatives are similar. Tests for the IIA assumption work by dropping one or more of the available options from the choice set and then reestimating the multinomial model. If the IIA assumption holds, then the estimates should not change very much. A statistical comparison of the two sets of estimates, one set from the model with a full set of alternatives, and the other from the model using a reduced set of alternatives, is carried out using a Hausman contrast test proposed by Hausman and McFadden (1984).[17]

EXAMPLE 16.12 | Postsecondary Education Multinomial Choice

The National Education Longitudinal Study of 1988 (NELS:88) was the first nationally representative longitudinal study of eighth-grade students in public and private schools in the United States. It was sponsored by the National Center for Education Statistics. In 1988, some 25,000 eighth-graders and their parents, teachers, and principals were surveyed. In 1990, these same students (who were then mostly 10th graders, and some dropouts) and their teachers and principals were surveyed again. In 1992, the second follow-up survey was conducted of students, mostly in the 12th grade, but dropouts, parents, teachers, school administrators, and high school transcripts were also surveyed. The third follow-up was in 1994, after most students had graduated.[18]

[17] "Specification Tests for the Multinomial Logit Model," *Econometrica*, 49, pp. 1219–1240. A brief explanation of the test may be found in Greene (2018, Chapter 18.2.4), op. cit., p. 767.

[18] The study and data are summarized in *National Education Longitudinal Study: 1988–1994, Descriptive Summary Report with an Essay on Access and Choice in Post-Secondary Education*, by Allen Sanderson, Bernard Dugoni, Kenneth Rasinski, and John Taylor, C. Dennis Carroll project officer, NCES 96-175, National Center for Education Statistics, March 1996.

We have taken a subset of the total data, namely those who stayed in the panel of data through the third follow-up. On this group, we have complete data on the individuals and their households, high-school grades, and test scores, as well as their postsecondary education choices. In the data file *nels_small*, we have 1000 observations on students who chose, upon graduating from high school, either no college (*PSECHOICE* = 1), a 2-year college (*PSECHOICE* = 2), or a 4-year college (*PSECHOICE* = 3). For illustration purposes, we focus on the explanatory variable *GRADES*, which is an index ranging from 1.0 (highest level, A+ grade) to 13.0 (lowest level, F grade) and represents combined performance in English, maths, and social studies.

Of the 1000 students, 22.2% selected not to attend a college upon graduation, 25.1% selected to attend a 2-year college, and 52.7% attended a 4-year college. The average value of *GRADES* is 6.53, with highest grade 1.74 and lowest grade 12.33. The estimated values of the parameters and their standard errors are given in Table 16.2. We selected the group who did not attend a college to be our base group, so that the parameters $\beta_{11} = \beta_{21} = 0$.

Based on these estimates, what can we say? Recall that a larger numerical value of *GRADES* represents a poorer academic performance. The parameter estimates for the coefficients of *GRADES* are negative and statistically significant. Using expression (16.22) on the effect of a change in an explanatory variable on the probability ratio, this means that if the value of *GRADES* increases, the probability that high-school graduates will choose a 2-year or a 4-year college goes down, relative to the probability of not attending college. This is the anticipated effect, as we expect that a poorer academic performance will increase the odds of not attending college.

We can also compute the estimated probability of each type of college choice using (16.19) for given values of *GRADES*. In our sample, the median value of *GRADES* is 6.64, and the top 5th percentile value is 2.635.[19] What are the choice probabilities of students with these grades? In Table 16.3, we show that the probability of choosing no college is 0.1810 for the student with median grades, but this probability is reduced to 0.0178 for students with top grades. Similarly, the probability of choosing a 2-year school is 0.2856 for the average student but is 0.0966 for the better student. Finally, the average student has a 0.5334 chance of selecting a 4-year college, but the better student has a 0.8857 chance of selecting a 4-year college.

The marginal effect of a change in *GRADES* on the choice probabilities can be calculated using (16.20). The marginal effect again depends on particular values for *GRADES*, and we report these in Table 16.3 for the median and 5th percentile students. An increase in *GRADES* of one point (worse performance) increases the probabilities of choosing either no college or a 2-year college and reduces the probability of attending a 4-year college. The probability of attending a 4-year college declines more for the average student than for the top student, given the one-point increase in *GRADES*. Note that for each value of *GRADES* the sum of the predicted probabilities is one, and the sum of the marginal effects is zero, except for rounding error. This is a feature of the multinomial logit specification.

TABLE 16.2 **Maximum Likelihood Estimates of PSE Choice**

Parameters	Estimates	Standard Errors	*t*-Statistics
β_{12}	2.5064	0.4183	5.99
β_{22}	−0.3088	0.0523	−5.91
β_{13}	5.7699	0.4043	14.27
β_{23}	−0.7062	0.0529	−13.34

TABLE 16.3 **Effects of Grades on Probability of PSE Choice**

PSE Choice	*GRADES*	$\hat{p}$	$\text{se}(\hat{p})$	Marginal Effect	se(ME)
No college	6.64	0.1810	0.0149	0.0841	0.0063
	2.635	0.0178	0.0047	0.0116	0.0022
Two-year college	6.64	0.2856	0.0161	0.0446	0.0076
	2.635	0.0966	0.0160	0.0335	0.0024
Four-year college	6.64	0.5334	0.0182	−0.1287	0.0095
	2.635	0.8857	0.0174	−0.0451	0.0030

[19] The 5th percentile value of *GRADES* is given as 2.635 which is halfway between observations 50 and 51 in this 1,000 observation data set. While this is a common way to calculate the 5th percentile, it is not the only way. Since $0.05 \times 1000 = 50$, some software will report the 50th value, after sorting according to increasing value, 2.63. Others may take a weighted average of the 50th and 51st values, such as $0.95 \times 2.63 + 0.05 \times 2.64 = 2.6305$. Thanks to Tom Doan (Estima) for noting this. Standard errors in Table 16.3 are computed via "the delta method," in a fashion similar to that described in Appendix 16A.

16.4 Conditional Logit

Suppose that a decision-maker must choose between several distinct alternatives, just as in the multinomial logit model. In a marketing context, suppose that our decision is between three types ($J = 3$) of soft drinks, say Pepsi, 7-Up, and Coke Classic, in 2-liter bottles. Shoppers will visit their supermarkets and make a choice, based on prices of the products and other factors. With the advent of supermarket scanners at checkout, data on purchases (what brand, how many units, and the price paid) are recorded. Of course, we also know the prices of the products that the consumer did not buy on a particular shopping occasion. The key point is that if we collect data on soda purchases from a variety of supermarkets, over a period of time, we observe consumer choices from the set of alternatives and we know the prices facing the shopper on each trip to the supermarket.

Let y_{i1}, y_{i2}, and y_{i3} be indicator variables that indicate the choice made by individual i. If alternative one (Pepsi) is selected, then $y_{i1} = 1$, $y_{i2} = 0$, and $y_{i3} = 0$. If alternative two (7-Up) is selected, then $y_{i1} = 0$, $y_{i2} = 1$, and $y_{i3} = 0$. If alternative 3 (Coke) is selected, then $y_{i1} = 0$, $y_{i2} = 0$, and $y_{i3} = 1$. The price facing individual i for brand j is $PRICE_{ij}$. That is, the price of Pepsi, 7-Up, and Coke is potentially different for each customer who purchases soda. Remember, different customers can shop at different supermarkets and at different times. Variables like *PRICE* are *individual- and alternative-specific* because they vary from individual to individual and are different for each choice the consumer might make. This type of information is very different from what we assumed was available in the multinomial logit model, where the explanatory variable x_i was *individual-specific*; it did not change across alternatives.

16.4.1 Conditional Logit Choice Probabilities

Our objective is to understand the factors that lead a consumer to choose one alternative over another. We construct a model for the probability that individual i chooses alternative j

$$p_{ij} = P[\text{individual } i \text{ chooses alternative } j]$$

The conditional logit model specifies these probabilities as

$$p_{ij} = \frac{\exp(\beta_{1j} + \beta_2 PRICE_{ij})}{\exp(\beta_{11} + \beta_2 PRICE_{i1}) + \exp(\beta_{12} + \beta_2 PRICE_{i2}) + \exp(\beta_{13} + \beta_2 PRICE_{i3})} \tag{16.23}$$

Note that unlike the probabilities for the multinomial logit model in (16.19), there is only one parameter β_2 relating the effect of each price to the choice probability p_{ij}. We have also included alternative specific constants (intercept terms). These cannot all be estimated, and one must be set to zero. We will set $\beta_{13} = 0$.

Estimation of the unknown parameters is by maximum likelihood. Suppose that we observe three individuals, who choose alternatives one, two, and three, respectively. Assuming that their choices are independent, then the probability of observing this outcome is

$$\begin{aligned} P(y_{11} = 1, y_{22} = 1, y_{33} = 1) &= p_{11} \times p_{22} \times p_{33} \\ &= \frac{\exp(\beta_{11} + \beta_2 PRICE_{11})}{\exp(\beta_{11} + \beta_2 PRICE_{11}) + \exp(\beta_{12} + \beta_2 PRICE_{12}) + \exp(\beta_2 PRICE_{13})} \\ &\quad \times \frac{\exp(\beta_{12} + \beta_2 PRICE_{22})}{\exp(\beta_{11} + \beta_2 PRICE_{21}) + \exp(\beta_{12} + \beta_2 PRICE_{22}) + \exp(\beta_2 PRICE_{23})} \\ &\quad \times \frac{\exp(\beta_2 PRICE_{33})}{\exp(\beta_{11} + \beta_2 PRICE_{31}) + \exp(\beta_{12} + \beta_2 PRICE_{32}) + \exp(\beta_2 PRICE_{33})} \\ &= L(\beta_{11}, \beta_{12}, \beta_2) \end{aligned}$$

16.4.2 Conditional Logit Postestimation Analysis

How a change in price affects the choice probability is different for "own price" changes and "cross-price" changes. Specifically, it can be shown that the own price effect is

$$\frac{\partial p_{ij}}{\partial PRICE_{ij}} = p_{ij}(1 - p_{ij})\beta_2 \tag{16.24}$$

The sign of β_2 indicates the direction of the own price effect.

The change in probability of alternative j being selected if the price of alternative k changes ($k \neq j$) is

$$\frac{\partial p_{ij}}{\partial PRICE_{ik}} = -p_{ij}p_{ik}\beta_2 \tag{16.25}$$

The cross-price effect is in the opposite direction of the own price effect.

An important feature of the conditional logit model is that the probability ratio between alternatives j and k is

$$\frac{p_{ij}}{p_{ik}} = \frac{\exp(\beta_{1j} + \beta_2 PRICE_{ij})}{\exp(\beta_{1k} + \beta_2 PRICE_{ik})} = \exp\left[(\beta_{1j} - \beta_{1k}) + \beta_2(PRICE_{ij} - PRICE_{ik})\right]$$

The probability ratio depends on the difference in prices but not on the prices themselves. As in the multinomial logit model, this ratio does not depend on the total number of alternatives, and there is the implicit assumption of the **independence of irrelevant alternatives (IIA)**. See the discussion at the end of Section 16.3.3. Models that do not require the IIA assumption have been developed, but they are difficult. These include the *multinomial probit* model, which is based on the normal distribution, and the *nested logit* and *mixed logit* models.[20]

EXAMPLE 16.13 | Conditional Logit Soft Drink Choice

We observe 1822 purchases, covering 104 weeks and 5 stores, in which a consumer purchased 2-liter bottles of either Pepsi (34.6%), 7-Up (37.4%), or Coke Classic (28%). These data are in the file *cola*. In the sample, the average price of Pepsi was \$1.23, 7-Up \$1.12, and Coke \$1.21. We estimate the conditional logit model shown in (16.22), and the estimates are shown in Table 16.4a.

TABLE 16.4a **Conditional Logit Parameter Estimates**

Variable	Estimate	Standard Error	*t*-Statistic	*p*-Value
PRICE(β_2)	−2.2964	0.1377	−16.68	0.000
PEPSI(β_{11})	0.2832	0.0624	4.54	0.000
7-*UP*(β_{12})	0.1038	0.0625	1.66	0.096

We see that all the parameter estimates are significantly different from zero at a 10% level of significance, and the sign of the coefficient of *PRICE* is negative. This means that a rise in the price of an individual brand will reduce the probability of its purchase, and the rise in the price of a competitive brand will increase the probability of its purchase. Table 16.4b contains the marginal effects of price changes on the probability of choosing Pepsi. The marginal effects are calculated using (16.24) and (16.25) with prices of Pepsi, 7-Up, and Coke set to \$1.00, \$1.25, and \$1.10, respectively. The standard errors are calculated using the delta method. Note two things about these estimates. First, they have the signs we anticipate. An increase in the price of Pepsi is estimated to have a negative effect on the probability of Pepsi purchase, while an increase in the price of either Coke or 7-Up increases the probability that Pepsi will be selected. Second, these values are very large for changes in probabilities because a "one-unit change" is \$1, which then represents almost a 100% change in price. For a 10-cent increase in

[20] For a brief description of these models at an advanced level, see William Greene, Econometric Analysis, Eighth Edition by (Upper Saddle River, NJ: Pearson Prentice Hall, 2018), Chapter 18.2.5. Mixed and nested logit models are important in applied research. David A. Hensher, John M. Rose, William H. Greene (2015) *Applied Choice Analysis, 2nd Edition*, Cambridge University Press, provide a comprehensive overview and integration of choice models, along with software instructions using the *NLOGIT* software package. Survey methodology is also discussed.

the prices the marginal effects, standard errors and interval estimate bounds should be multiplied by 0.10.

TABLE 16.4b **Marginal Effect of Price on Probability of Pepsi Choice**

PRICE	Marginal Effect	Standard Error	95% Interval Estimate
COKE	0.3211	0.0254	[0.2712, 0.3709]
PEPSI	−0.5734	0.0350	[−0.6421, −0.5048]
7-UP	0.2524	0.0142	[0.2246, 0.2802]

As an alternative to computing marginal effects, we can compute specific probabilities at given values of the explanatory variables. For example, at the prices used for Table 16.4b, the estimated probability of selecting Pepsi is then

$$\hat{p}_{i1} = \frac{\exp(\tilde{\beta}_{11} + \tilde{\beta}_2 \times 1.00)}{\left[\exp(\tilde{\beta}_{11} + \tilde{\beta}_2 \times 1.00) + \exp(\tilde{\beta}_{12} + \tilde{\beta}_2 \times 1.25) + \exp(\tilde{\beta}_2 \times 1.10)\right]} = 0.4832$$

The standard error for this predicted probability is 0.0154, which is computed via the delta method. If we raise the price of Pepsi to \$1.10, we estimate that the probability of its purchase falls to 0.4263 (se = 0.0135). If the price of Pepsi stays at \$1.00 but we increase the price of Coke by 15 cents, then we estimate that the probability of a consumer selecting Pepsi rises by 0.0445 (se = 0.0033). These numbers indicate to us the responsiveness of brand choice to changes in prices, much like elasticities.

16.5 Ordered Choice Models

The choice options in multinomial and conditional logit models have no natural ordering or arrangement. However, in some cases, choices are ordered in a specific way. Examples include the following:

1. Results of opinion surveys in which responses can be strongly in disagreement, in disagreement, neutral, in agreement, or strongly in agreement.
2. Assignment of grades or work performance ratings. Students receive grades A, B, C, D, and F, which are ordered on the basis of a teacher's evaluation of their performance. Employees are often given evaluations on scales such as outstanding, very good, good, fair, and poor, which are similar in spirit.
3. Standard and Poor's rates bonds as AAA, AA, A, BBB, and so on, as a judgment about the credit worthiness of the company or country issuing a bond, and how risky the investment might be.
4. Levels of employment as unemployed, part time, or full time.

When modeling these types of outcomes, numerical values are assigned to the outcomes, but the numerical values are **ordinal** and reflect only the ranking of the outcomes. In the first example, we might assign a dependent variable y the values

$$y = \begin{cases} 1 & \text{strongly disagree} \\ 2 & \text{disagree} \\ 3 & \text{neutral} \\ 4 & \text{agree} \\ 5 & \text{strongly agree} \end{cases}$$

In Section 16.3, we considered the problem of choosing what type of college to attend after graduating from high school as an illustration of a choice among unordered alternatives.

However, in this particular case, there may in fact be natural ordering. We might rank the possibilities as

$$y = \begin{cases} 3 & \text{four-year college (the full college experience)} \\ 2 & \text{two-year college (a partial college experience)} \\ 1 & \text{no college} \end{cases} \tag{16.26}$$

The usual linear regression model is not appropriate for such data, because in regression we would treat the y-values as having some numerical meaning when they do not. In the following section, we discuss how probabilities of each choice might be modeled.

16.5.1 Ordinal Probit Choice Probabilities

When faced with a ranking problem, we develop a "sentiment" about how we feel concerning the alternative choices, and the higher the sentiment, the more likely a higher ranked alternative will be chosen. This sentiment is, of course, unobservable to the econometrician. Unobservable variables that enter decisions are called **latent variables**, and we will denote our sentiment toward the ranked alternatives by y_i^*, with the "star" reminding us that this variable is unobserved. See Appendix 16B for more on latent variables.

Microeconomics is well described as the "science of choice." Economic theory will suggest that certain factors (observable variables) may affect how we feel about the alternatives facing us. As a concrete example, let us think about what factors might lead a high-school graduate to choose among the alternatives "no college," "2-year college," and "4-year college" as described by the ordered choices in (16.26). Some factors that affect this choice are household income, the student's high-school grades, how close a 2- or 4-year college is to the home, whether parents had attended a 4-year college, and so on. For simplicity, let us focus on the single explanatory variable *GRADES*. The model is then

$$y_i^* = \beta GRADES_i + e_i$$

This model is not a regression model because the dependent variable is unobservable. Consequently, it is sometimes called an **index model**. The error term is present for the usual reasons. The choices we observe are based on a comparison of "sentiment" toward higher education y_i^* relative to certain thresholds, as shown in Figure 16.2.

Because there are $M = 3$ alternatives, there are $M - 1 = 2$ thresholds μ_1 and μ_2, with $\mu_1 < \mu_2$. The index model does not contain an intercept because it would be exactly collinear with the threshold variables. If sentiment toward higher education is in the lowest category, then $y_i^* \le \mu_1$ and the alternative "no college" is chosen, if $\mu_1 < y_i^* \le \mu_2$ then the alternative "2-year college"

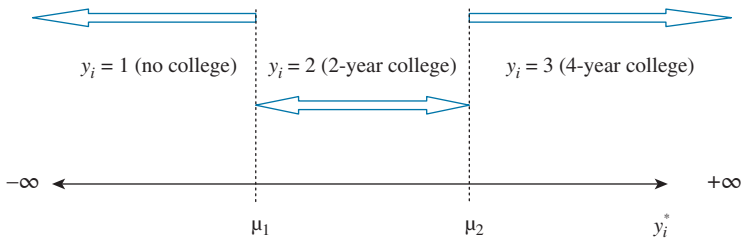


FIGURE 16.2 **Ordinal choices relative to thresholds.**

is chosen, and if sentiment toward higher education is in the highest category, then $y_i^* > \mu_2$ and "4-year college" is chosen. That is,

$$y_i = \begin{cases} 3\,(\text{four-year college}) & \text{if } y_i^* > \mu_2 \\ 2\,(\text{two-year college}) & \text{if } \mu_1 < y_i^* \le \mu_2 \\ 1\,(\text{no college}) & \text{if } y_i^* \le \mu_1 \end{cases}$$

We are able to represent the probabilities of these outcomes if we assume a particular probability distribution for y_i^*, or equivalently for the random error e_i. If we assume that the errors have the standard normal distribution, $N(0, 1)$, an assumption that defines the **ordered probit model**, then we can calculate the following:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \le \mu_1) = P(\beta GRADES_i + e_i \le \mu_1) \\ &= P(e_i \le \mu_1 - \beta GRADES_i) \\ &= \Phi(\mu_1 - \beta GRADES_i) \end{aligned}$$

$$\begin{aligned} P(y_i = 2) &= P(\mu_1 < y_i^* \le \mu_2) = P(\mu_1 < \beta GRADES_i + e_i \le \mu_2) \\ &= P(\mu_1 - \beta GRADES_i < e_i \le \mu_2 - \beta GRADES_i) \\ &= \Phi(\mu_2 - \beta GRADES_i) - \Phi(\mu_1 - \beta GRADES_i) \end{aligned}$$

and the probability that $y = 3$ is

$$\begin{aligned} P(y_i = 3) &= P(y_i^* > \mu_2) = P(\beta GRADES_i + e_i > \mu_2) \\ &= P(e_i > \mu_2 - \beta GRADES_i) \\ &= 1 - \Phi(\mu_2 - \beta GRADES_i) \end{aligned}$$

16.5.2 Ordered Probit Estimation and Interpretation

Estimation, as with previous choice models, is by maximum likelihood. If we observe a random sample of $N = 3$ individuals, with the first not going to college ($y_1 = 1$), the second attending a 2-year college ($y_2 = 2$), and the third attending a 4-year college ($y_3 = 3$), then the likelihood function is

$$L(\beta, \mu_1, \mu_2) = P(y_1 = 1) \times P(y_2 = 2) \times P(y_3 = 3)$$

Note that the probabilities depend on the unknown parameters μ_1 and μ_2 as well as the index function parameter β. These parameters are obtained by maximizing the log-likelihood function using numerical methods. Econometric software includes options for both **ordered probit**, which depends on the errors being standard normal, and **ordered logit**, which depends on the assumption that the random errors follow a logistic distribution. Most economists will use the normality assumption, but many other social scientists use the logistic. In the end, there is little difference between the results.

The types of questions we can answer with this model are the following:

1. What is the probability that a high-school graduate with $GRADES = 2.5$ (on a 13-point scale, with one being the highest) will attend a 2-year college? The answer is obtained by plugging in the specific value of *GRADES* into the estimated probability using the maximum likelihood estimates of the parameters,

$$\hat{P}(y = 2|GRADES = 2.5) = \Phi(\tilde{\mu}_2 - \tilde{\beta} \times 2.5) - \Phi(\tilde{\mu}_1 - \tilde{\beta} \times 2.5)$$

2. What is the difference in probability of attending a 4-year college for two students, one with $GRADES = 2.5$ and another with $GRADES = 4.5$? The difference in the probabilities is calculated directly as

$$\hat{P}(y = 3|GRADES = 4.5) - \hat{P}(y = 3|GRADES = 2.5)$$

3. If we treat *GRADES* as a continuous variable, what is the marginal effect on the probability of each outcome, given a one-unit change in *GRADES*? These derivatives are

$$\frac{\partial P(y = 1)}{\partial GRADES} = -\phi(\mu_1 - \beta GRADES) \times \beta$$

$$\frac{\partial P(y = 2)}{\partial GRADES} = \left[\phi(\mu_1 - \beta GRADES) - \phi(\mu_2 - \beta GRADES)\right] \times \beta$$

$$\frac{\partial P(y = 3)}{\partial GRADES} = \phi(\mu_2 - \beta GRADES) \times \beta$$

In these expressions, "$\phi(\cdot)$" denotes the *pdf* of a standard normal distribution, and its values are always positive. Consequently, the sign of the parameter β is opposite the direction of the marginal effect for the lowest category, but it indicates the direction of the marginal effect for the highest category. The direction of the marginal effect for the middle category goes one way or the other, depending on the sign of the difference in brackets.

There are a variety of other devices that can be used to analyze the outcomes, including some useful graphics. For more on these, see (from a social science perspective) *Regression Models for Categorical and Limited Dependent Variables* by J. Scott Long (Sage Publications, 1997, Chapter 5) or (from a marketing perspective) *Quantitative Models in Marketing Research* by Philip Hans Franses and Richard Paap (Cambridge University Press, 2001, Chapter 6). A comprehensive reference is by William H. Greene and David A. Hensher (2010) *Modeling Ordered Choices: A Primer*, Cambridge University Press.

EXAMPLE 16.14 | Postsecondary Education Ordered Choice Model

To illustrate, we use the college choice data introduced in Section 16.3 and contained in the data file *nels_small*. We treat *PSECHOICE* as an ordered variable with 1 representing the least favored alternative (no college) and 3 denoting the most favored alternative (4-year college). The estimation results are in Table 16.5.

TABLE 16.5 **Ordered Probit Parameter Estimates for PSE Choice**

Parameters	Estimates	Standard Errors
β	−0.3066	0.0191
μ_1	−2.9456	0.1468
μ_2	−2.0900	0.1358

The estimated coefficient of *GRADES* is negative, indicating that the probability of attending a 4-year college goes down when *GRADES* increase (indicating a worse performance), and the probability of the lowest ranked choice, attending no college, increases. Let us examine the marginal effects of an increase in *GRADES* on attending a 4-year college. For a student with median grades (6.64), the marginal effect is −0.1221, and for a student in the 5th percentile (2.635), the marginal effect is −0.0538. These are similar in magnitude to the marginal effects shown in Table 16.3.

16.6 Models for Count Data

When the dependent variable in a regression model is a count of the number of occurrences of an event, the outcome variable is $y = 0, 1, 2, 3, \ldots$ These numbers are actual counts and thus different from the ordinal numbers of the previous section. Examples include the following:

- The number of trips to a physician a person makes during a year.
- The number of fishing trips taken by a person during the previous year.
- The number of children in a household.
- The number of automobile accidents at a particular intersection during a month.
- The number of televisions in a household.
- The number of alcoholic drinks a college student takes in a week.

While we are again interested in explaining and estimating probabilities, such as the probability that an individual will take two or more trips to the doctor during a year, the probability distribution we use as a foundation is the Poisson, not the normal or the logistic. If Y is a **Poisson random variable**, then its probability function is

$$f(y) = P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots \tag{16.27}$$

The factorial (!) term $y! = y \times (y-1) \times (y-2) \times \cdots \times 1$. This probability function has one parameter, λ, which is the mean (and variance) of Y. That is, $E(Y) = \text{var}(Y) = \lambda$. In a regression model, we try to explain the behavior of $E(Y)$ as a function of some explanatory variables. We do the same here, keeping the value of $E(Y) \geq 0$ by defining

$$E(Y|x) = \lambda = \exp(\beta_1 + \beta_2 x) \tag{16.28}$$

This choice defines the **Poisson regression model** for count data.

16.6.1 Maximum Likelihood Estimation of the Poisson Regression Model

The parameters β_1 and β_2 in (16.28) can be estimated by maximum likelihood. Suppose that we randomly select $N = 3$ individuals from a population and observe that their counts are $y_1 = 0$, $y_2 = 2$, and $y_3 = 2$, indicating 0, 2, and 2 occurrences of the event for these three individuals. Recall that the likelihood function is the joint probability function of the observed data, interpreted as a function of the unknown parameters. That is,

$$L(\beta_1, \beta_2) = P(Y = 0) \times P(Y = 2) \times P(Y = 2)$$

This product of functions like (16.27) will be very complicated and difficult to maximize. However, in practice, maximum likelihood estimation is carried out by maximizing the logarithm of the likelihood function, or

$$\ln L(\beta_1, \beta_2) = \ln P(Y = 0) + \ln P(Y = 2) + \ln P(Y = 2)$$

Using (16.28) for λ, the log of the probability function is

$$\begin{aligned} \ln\left[P(Y = y|x)\right] &= \ln\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right] = -\lambda + y\ln(\lambda) - \ln(y!) \\ &= -\exp(\beta_1 + \beta_2 x) + \left[y \times (\beta_1 + \beta_2 x)\right] - \ln(y!) \end{aligned}$$

Then the log-likelihood function, given a sample of N observations, becomes

$$\ln L(\beta_1, \beta_2) = \sum_{i=1}^{N} \left\{ -\exp(\beta_1 + \beta_2 x_i) + y_i \times (\beta_1 + \beta_2 x_i) - \ln(y_i!) \right\}$$

This log-likelihood function is a function of only β_1 and β_2 once we substitute in the data values (y_i, x_i). The log-likelihood function itself is still a nonlinear function of the unknown parameters, and the maximum likelihood estimates must be obtained by numerical methods. Econometric software has options that allow for the maximum likelihood estimation of count models with the click of a button.

16.6.2 Interpreting the Poisson Regression Model

As in other modeling situations, we would like to use the estimated model to predict outcomes, determine the marginal effect of a change in an explanatory variable on the mean of the dependent variable, and test the significance of coefficients.

Estimation of the conditional mean of y is straightforward. Given the maximum likelihood estimates $\tilde{\beta}_1$ and $\tilde{\beta}_2$, and given a value of the explanatory variable x_0,

$$\widehat{E(y_0)} = \tilde{\lambda}_0 = \exp(\tilde{\beta}_1 + \tilde{\beta}_2 x_0)$$

This value is an estimate of the expected number of occurrences observed if x takes the value x_0. The probability of a particular number of occurrences can be estimated by inserting the estimated conditional mean into the probability function, as

$$\widehat{P(Y = y)} = \frac{\exp(-\tilde{\lambda}_0)\,\tilde{\lambda}_0^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

The marginal effect of a change in a continuous variable x in the Poisson regression model is not simply given by the parameter because the conditional mean model is a nonlinear function of the parameters. Using our specification that the conditional mean is given by $E(y_i|x_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$, and using rules for derivatives of exponential functions, we obtain the marginal effect

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \lambda_i \beta_2 \tag{16.29}$$

To estimate this marginal effect, replace the parameters by their maximum likelihood estimates and select a value for x. The marginal effect is different depending on the value of x chosen. A useful fact about the Poisson model is that the conditional mean $E(y_i|x_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$ is always positive because the exponential function is always positive. Thus, the direction of the marginal effect can be determined from the sign of the coefficient β_2.

Equation (16.29) can be expressed as a percentage, which can be useful:

$$\frac{\%\Delta E(y_i|\mathbf{x})}{\Delta x_i} = 100 \frac{\partial E(y_i|\mathbf{x}) / E(y_i|\mathbf{x})}{\partial x_i} = 100\beta_2 \%$$

If x is not transformed, then a one-unit change in x leads to $100\beta_2\%$ change in the conditional mean.

Suppose that the conditional mean function contains a indicator variable, how do we calculate its effect? If $E(y_i|\mathbf{x}) = \lambda_i = \exp(\beta_1 + \beta_2 x_i + \delta D_i)$, we can examine the conditional expectation when $D = 0$ and when $D = 1$.

$$E(y_i|x_i, D_i = 0) = \exp(\beta_1 + \beta_2 x_i)$$

$$E(y_i|x_i, D_i = 1) = \exp(\beta_1 + \beta_2 x_i + \delta)$$

Then, the percentage change in the conditional mean is

$$100\left[\frac{\exp(\beta_1+\beta_2 x_i+\delta)-\exp(\beta_1+\beta_2 x_i)}{\exp(\beta_1+\beta_2 x_i)}\right]\% = 100\left[e^{\delta}-1\right]\%$$

This is identical to the expression we obtained for the effect of an indicator variable in a log-linear model. See Section 7.3.

Finally, hypothesis testing can be carried out using standard methods. The maximum likelihood estimators are asymptotically normal with a variance of a known form. The actual expression for the variance is complicated and involves matrix expressions, so we will not report the formula here.[21] Econometric software has the variance expressions encoded, and along with parameter estimates, it will provide standard errors, t-statistics, and p-values, which are used as always.

EXAMPLE 16.15 | A Count Model for the Number of Doctor Visits

The economic analysis of the health care system is a vital area of research and public interest. In this example, we consider data used by Riphahn, Wambach, and Million (2003).[22] The data file *rwm88_small* contains data on 1,200 individuals' number of doctor visits in the past three months (*DOCVIS*), their age in years (*AGE*), their sex (*FEMALE*), and whether or not they had public insurance (*PUBLIC*). The frequencies of doctor visits are illustrated in Table 16.6, with 90.5% of the sample having eight or fewer visits.

TABLE 16.6 **Number of Doctor Visits (*DOCVIS*)**

DOCVIS	Number
0	443
1	200
2	163
3	111
4	51
5	49
6	37
7	7
8	25

These are numerical **count data** (number of times an event occurs) so that the Poisson model is a feasible choice. Applying maximum likelihood estimation, we obtain the fitted model

$$\underset{(\text{se})}{\widehat{E(DOCVIS)}} = \exp\underset{(0.0918)}{(-0.0030}+\underset{(0.0015)}{0.0116}AGE+\underset{(0.0335)}{0.1283}FEMALE+\underset{(0.0680)}{0.5726}PUBLIC)$$

What can we say about these results? First, the coefficient estimates are all positive, implying that older individuals, females and those with public health insurance will have more doctor visits. Second, the coefficients of *AGE*, *FEMALE* and *PUBLIC* are significantly different from zero, with p-values less than 0.01. Using the fitted model, we can estimate the expected number of doctor visits. For example, the first person in the sample is a 29-year-old female who has public insurance. Substituting these values we estimate her expected number of doctor visits to be 2.816, or 3.0 rounded to the nearest integer. Her actual number of doctor visits was zero.

Using the notion of generalized-R^2, we can get a notion of how well the model fits the data by computing the squared correlation between *DOCVIS* and the predicted number of visits. If we use the rounded values, for example, 3.0 instead of 3.33, the correlation is 0.1179 giving $R_g^2 = (0.1179)^2 = 0.0139$. The fit for this simple model is not very good as we might well expect. This model does not account for so many important factors, such as income, general health status, and so on. Different software packages report many different values, sometimes called *pseudo*-R^2,

[21] See J. Scott Long, *Regression Models for Categorical and Limited Dependent Variables* (Thousand Oaks, CA: Sage Publications, 1997), Chapter 8. A much more advanced and specialized reference is *Regression Analysis of Count Data* Second Edition by A. Colin Cameron and Pravin K. Trivedi (Cambridge, UK: Cambridge University Press, 2013).

[22] Regina T. Riphahn, Achim Wambach, and Andreas Million, "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation", *Journal of Applied Econometrics*, Vol. 18, No. 4, 2003, pp. 387–405.

with different meanings as well. We urge you to ignore all these values, including R_g^2.

Instead of an R^2-like number, it is a good idea to report a test of overall model significance, analogous to the overall F-test for the regression model. The null hypothesis is that all the model coefficients, except the intercept, are equal to zero. We recommend the likelihood ratio statistic. See Section 16.2.7 for a discussion of this test in the context of the probit model. The test statistic is $LR = 2(\ln L_U - \ln L_R)$ where $\ln L_U$ is the value of the log-likelihood function for the full and unrestricted model and $\ln L_R$ is the value of the log-likelihood function for the restricted model that assumes that the hypothesis is true. The restricted model in this case is $E(DOCVIS) = \exp(\gamma_1)$. If the null hypothesis is true, the LR test statistic has a $\chi^2_{(3)}$-distribution in large samples. In our example, $LR = 174.93$ and the 0.95 percentile of the $\chi^2_{(3)}$-distribution is 7.815. We reject the null hypothesis at the 5% level of significance, and we conclude that at least one variable makes a significant impact on the number of doctor visits.

What about the magnitudes of the effects of these variables on the number of doctor visits? Treating *AGE* as continuous we can use (16.29) to compute a marginal effect,

$$\widehat{\frac{\partial E(DOCVIS)}{\partial AGE}} = \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE + 0.5726PUBLIC) \times 0.0116$$

To evaluate this effect, we must insert values for *AGE*, *FEMALE,* and *PUBLIC*. Let *FEMALE* = 1 and *PUBLIC* = 1. If *AGE* = 30, the estimate is 0.0332, with the 95% interval estimate being [0.0261, 0.0402]. That is, we estimate for a 30-year-old female with public insurance an additional year of age will increase her expected number of doctor visits in a 3-month period by 0.0332. Because the marginal effect is a nonlinear function of the estimated parameters, the interval estimate uses a standard error calculated using the delta method. For *AGE* = 70, it is 0.0528 [0.0355, 0.0702]. The effect of another year of age is greater for older individuals, as you would expect.

Both *FEMALE* and *PUBLIC* are indicator variables, taking values zero and one. For these variables, we cannot evaluate the "marginal effect" using a derivative. Instead, we estimate the difference between the expected number of doctor visits for the two cases. For example,

$$\Delta E(DOCVIS) = E(DOCVIS|PUBLIC = 1) - E(DOCVIS|PUBLIC = 0)$$

The calculated value of the difference is

$$\widehat{\Delta E(DOCVIS)} = \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE + 0.5726) - \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE)$$

We estimate the difference for a 30-year-old female to be 1.24 [1.00, 1.48], and for a 70-year-old female, it is 1.98 [1.59, 2.36]. Women with public insurance visit the doctor significantly more than women of the same age who do not have public insurance.

There are many generalizations of the Poisson model that are used in applied work. One generalization is called the *negative binomial model*. It can be used when an assumption implicit in the Poisson model is violated, namely that the variance of Poisson variable is equal to its expected value, that is $\text{var}(Y) = E(Y)$ for Poisson random variables. There are tests for whether this assumption holds. These are sometimes called *tests for overdispersion*. A second type of possible misspecification is illustrated by the following question: How many extramarital affairs did you have in the last year?[23] The first thing to note is that the question is relevant only for married individuals. The possible answers are zero, one, two, three, etc. However, here a "zero" might mean two different things. It might mean, "I would *never* cheat on my spouse!!" or it might mean, "Well, I have cheated in the past, but not in the last year." Statistically, in this situation, there will be "too many zeros" for the standard Poisson distribution. As a result, there are some *zero-inflated* versions of the Poisson model (*ZIP*) that may be a better choice. These extensions of the Poisson model are quite fascinating and useful but beyond the scope of this book.[24]

[23] Ray Fair (1978) "The Theory of Extramarital Affairs," *Journal of Political Economy*, 86(1), 45–61.

[24] Two excellent but advanced references are: A. Colin Cameron and Pravin K. Trivedi (2013) *Regression Analysis of Count Data*, Cambridge University Press; and Rainer Winkelman (2008) *Econometric Analysis of Count Data*, Springer.

16.7 Limited Dependent Variables

In the previous sections of this chapter, we reviewed choice behavior models that have dependent variables that are discrete variables. When a model has a discrete dependent variable, the usual regression methods we have studied must be modified. In this section, we present another case in which standard least squares estimation of a regression model fails.

16.7.1 Maximum Likelihood Estimation of the Simple Linear Regression Model

We have stressed the least squares and method of moments estimators when estimating the simple linear regression model. Another option is maximum likelihood estimation (MLE). Our discussion of the method will be in the context of the simple linear model with one explanatory variable, but the method extends to the case of multiple regression with more explanatory variables. We discuss this now because several strategies for estimating **limited dependent variable** models are tied to MLE. In this case, we make assumptions SR1–SR6, which include the assumption about the conditional normality of the random errors. When the assumption of conditionally normal errors is made, we write $e_i|x_i \sim N(0, \sigma^2)$, and also then $y_i|x_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$. It is a very strong assumption when it is made, and it is not necessary for least squares estimation, so we have called it an *optional* assumption. For maximum likelihood estimation, it is *not* optional. It is necessary to assume a distribution for the data so that we can form the likelihood function.

If the data (y_i, x_i) pairs are drawn independently, then the conditional joint *pdf* of the data is

$$f(y_1, y_2, \ldots, y_N|\mathbf{x}, \beta_1, \beta_2, \sigma^2) = f(y_1|x_1, \beta_1, \beta_2, \sigma^2) \times \cdots \times f(y_N|x_N, \beta_1, \beta_2, \sigma^2) \quad (16.30a)$$

where

$$f(y_i|x_i, \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}\right) \quad (16.30b)$$

Writing out the product we have

$$f(y_1, \ldots, y_N|\mathbf{x}, \beta_1, \beta_2, \sigma^2) = (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right]$$

$$= L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) \quad (16.30c)$$

The likelihood function $L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x})$ is the joint *pdf* interpreted as function of the unknown parameters, conditional on the data. In practice, we maximize the log-likelihood,

$$\ln L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) = -(N/2)\ln(2\pi) - (N/2)\ln(\sigma^2)$$

$$- \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2 \quad (16.30d)$$

This looks quite intimidating to maximize, but this is one of the times we can actually maximize the log-likelihood using calculus. See Exercise 16.1 for hints. The maximum likelihood estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the OLS estimators, which have all their usual properties including a conditionally normal distribution. The MLE of the error variance is $\tilde{\sigma}^2 = \sum \hat{e}_i^2/N$, which is the sum of the squared least squares residuals divided by the sample size, with no degrees of freedom correction. This estimator is consistent but biased.

16.7.2 Truncated Regression

The first limited dependent variable model we consider is **truncated regression**. In this model, we only observe the data (y_i, x_i) when $y_i > 0$. How can this happen? Imagine collecting survey data by waiting at the checkout station in a supermarket. As customers exit you ask "How much did you spend today?" The answer will be some positive number given that they have just paid for their purchases. If the random error is conditionally normal, then the *pdf* of $(y_i|y_i > 0, x_i)$ is *truncated normal*. The properties of the truncated normal distribution are discussed in Appendix B.3.5. In this case, the truncated normal density function is

$$f(y_i|y_i > 0, x_i, \beta_1, \beta_2, \sigma^2) = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{P(y_i > 0|x_i, \beta_1, \beta_2, \sigma^2)} = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{\Phi\left(\frac{\beta_1 + \beta_2 x_i}{\sigma}\right)} = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{\Phi_i} \tag{16.31}$$

Here we use $\Phi_i = \Phi[(\beta_1 + \beta_2 x_i)/\sigma]$ as a simplifying notation. See Exercise 16.2 for hints on obtaining (16.31). The log-likelihood function is

$$\ln L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) = -\sum_{i=1}^{N} \ln \Phi_i - (N/2)\ln(2\pi) - (N/2)\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2 \tag{16.32}$$

Maximization of this log-likelihood has to be done using numerical methods, but econometric software has simple commands to obtain the estimates of β_1, β_2, and σ^2. The question then becomes, what can we do with these estimates? The answer is, all the usual things. For the calculation of marginal effects, use the conditional mean function

$$E(y_i|y_i > 0, x_i) = \beta_1 + \beta_2 x_i + \sigma\frac{\phi[(\beta_1 + \beta_2 x_i)/\sigma]}{\Phi[(\beta_1 + \beta_2 x_i)/\sigma]} = \beta_1 + \beta_2 x_i + \sigma\lambda(\alpha_i) \tag{16.33}$$

where $\lambda(\alpha_i)$ is the inverse Mill's ratio (IMR) mentioned in Appendix B.3.5 and $\alpha_i = (\beta_1 + \beta_2 x_i)/\sigma$. This is a bit of a mess isn't it? If x_i is continuous, the marginal effect is the derivative of this expression, $dE(y_i|y_i > 0, x_i)/dx_i = \beta_2(1 - \delta_i)$, where $\delta_i = \lambda(\alpha_i)[\lambda(\alpha_i) - \alpha_i]$, which is even more messy.[25] Because $0 < \delta_i < 1$, the marginal effect is only a fraction of the parameter value. Once again econometricians in conjunction with computer programmers have made our lives much easier than would otherwise be true and these quantities can be calculated.

16.7.3 Censored Samples and Regression

Censored samples are similar to truncated samples but have more information. In a *truncated sample*, we observe (y_i, x_i) when $y_i > 0$. For censored samples, we observe x_i for all individuals, but the outcome values are of two different types. In a survey of households, suppose we ask "How much did you spend on major appliances, such as refrigerators or washing machines, last month?" For many households, the answer will be \$0, as they made no such purchases in the previous month. For others, the answer will be a positive value, if such a purchase was made. This is the outcome variable, y_i. On the other hand, the survey will include income and other characteristics of the household, which are explanatory variables, x_i. This is called a *censored sample*, with a substantial fraction of the observations taking a *limit value*, in this case \$0. We are

[25] See Greene (2018), page 932–933.

interested in estimating the relationship between expenditures and an x_i. What should we do? There are a number of strategies. We will mention four, two that work and two that do not work.

Strategy 1 **Delete the limit observations and apply OLS**

A simple strategy is to drop from the sample the observations with $y_i = 0$ and go ahead. This strategy does not work. The usual OLS model, for $y_i > 0$, is $y_i = \beta_1 + \beta_2 x_i + u_i$, where u_i is an error term. We usually think of this model as resulting from the sum of a systematic part, the regression function, and a random error. That is, $y_i = E(y_i|y_i > 0, x_i) + e_i$. The conditional mean function is given by (16.33), so that

$$\begin{aligned} y_i = E(y_i|y_i > 0, x_i) + e_i &= \beta_1 + \beta_2 x_i + \sigma \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} + e_i \\ &= \beta_1 + \beta_2 x_i + \left\{ \sigma \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} + e_i \right\} \\ &= \beta_1 + \beta_2 x_i + u_i \end{aligned} \tag{16.34}$$

The error term u_i is not simple. It consists of the random component e_i and a complicated function of x_i. The error term u_i will be correlated with x_i, making OLS biased and inconsistent, which is not the result we want.

Strategy 2 **Retain all observations and apply OLS**

This strategy does not work. Using the definition for conditional expectation,

$$\begin{aligned} E(y_i|x_i) &= E(y_i|y_i > 0, x_i) \times P(y_i > 0) + E(y_i|y_i = 0, x_i) \times P(y_i = 0) \\ &= E(y_i|y_i > 0, x_i) \times \left\{1 - \Phi\left[-(\beta_1 + \beta_2 x_i)/\sigma\right]\right\} \\ &= E(y_i|y_i > 0, x_i) \times \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right] \\ &= \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]\beta_1 + \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]\beta_2 x_i + \sigma\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right] \end{aligned}$$

Simply estimating $y_i = \beta_1 + \beta_2 x_i + u_i$ by OLS clearly is inappropriate.

Strategy 3 **Heckman's two-step estimator**

The problem with Strategy 1 is that the error term u_i includes two components and one of them is correlated with the variable x_i. This is analogous to an omitted variables problem, the solution of which is to not omit the variable, but include it in the regression. That is, we would like to estimate the model

$$y_i = \beta_1 + \beta_2 x_i + \sigma\lambda_i + e_i$$

where

$$\lambda_i = \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} = \frac{\phi(\beta_1^* + \beta_2^* x_i)}{\Phi(\beta_1^* + \beta_2^* x_i)}$$

where $\beta_1^* = \beta_1/\sigma$ and $\beta_2^* = \beta_2/\sigma$. Nobel Prize winner James Heckman realized that while λ_i is unknown it can be consistently estimated as $\tilde{\lambda}_i = \phi(\tilde{\beta}_1^* + \tilde{\beta}_2^* x_i)/\Phi(\tilde{\beta}_1^* + \tilde{\beta}_2^* x_i)$ where $\tilde{\beta}_1^*$ and $\tilde{\beta}_2^*$ come from a probit model with dependent variable $d_i = 1$ if $y_i > 0$, and $d_i = 0$ if $y_i = 0$, and with explanatory variable x_i. Then the model we estimate by OLS is

$$y_i = \beta_1 + \beta_2 x_i + \sigma\tilde{\lambda}_i + e_i^*$$

It is called a two-step estimator because we use estimates from a first step, probit, and then a second step, OLS. The estimator is consistent and while correct standard errors are complicated, they are known and can be obtained.

Strategy 4 Maximum likelihood estimation: Tobit

Heckman's two-step estimator is consistent but not efficient. There is a maximum likelihood estimation procedure that is preferable. It is called **Tobit** in honor of James Tobin, winner of the 1981 Nobel Prize in Economics, who first studied the model.

Tobit is a maximum likelihood procedure that recognizes that we have data of two sorts, the limit observations ($y = 0$) and the nonlimit observations ($y > 0$). The two types of observations that we observe, the limit observations and those that are positive, are generated by a latent variable y^* crossing the zero threshold or not crossing that threshold. The (**probit**) probability that $y = 0$ is

$$P(y = 0|\mathbf{x}) = P(y^* \le 0|\mathbf{x}) = 1 - \Phi\left[(\beta_1 + \beta_2 x)/\sigma\right]$$

If we observe a positive value of y_i, then the term that enters the likelihood function is the normal *pdf* with mean $\beta_1 + \beta_2 x_i$ and variance σ^2. The full likelihood function is the product of the probabilities that the limit observations occur times the *pdfs* for all the positive, nonlimit, observations. Using "large pi" notation to denote multiplication, the likelihood function is

$$L(\beta_1, \beta_2, \sigma|\mathbf{x}, \mathbf{y}) = \prod_{y_i=0}\left\{1 - \Phi\left(\frac{\beta_1 + \beta_2 x_i}{\sigma}\right)\right\} \times \prod_{y_i>0}\left\{(2\pi\sigma^2)^{-0.5}\exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2\right)\right\}$$

This complicated-looking likelihood function is maximized numerically using econometric software.[26] The maximum likelihood estimator is consistent and asymptotically normal, with a known covariance matrix.[27]

16.7.4 Tobit Model Interpretation

In the **Tobit model**, the parameters β_1 and β_2 are the intercept and slope of the latent variable model (16.31). In practice, we are interested in the marginal effect of a change in x on either the regression function of the observed data $E(y|x)$ or the regression function conditional on $y > 0$, $E(y|x, y > 0)$. As we indicated earlier, these functions are not straight lines. Their graphs are shown in Figure 16.3. The slope of each changes at each value of x. The slope of $E(y|x)$ has a relatively simple form, being a scale factor times the parameter value; it is

$$\frac{\partial E(y|x)}{\partial x} = \beta_2 \Phi\left(\frac{\beta_1 + \beta_2 x}{\sigma}\right) \tag{16.35}$$

where Φ is the cumulative distribution function (*cdf*) of the standard normal random variable that is evaluated at the estimates and a particular x-value. Because the *cdf* values are positive, the sign of the coefficient tells the direction of the marginal effect, but the magnitude of the marginal effect depends on both the coefficient and the *cdf*. If $\beta_2 > 0$, as x increases, the *cdf* function approaches one, and the slope of the regression function approaches that of the latent variable

[26] Tobit requires data on both the limit values of $y = 0$ and also the nonlimit values for which $y > 0$. Sometimes, it is possible that we do not observe the limit values; in such a case, the sample is said to be truncated. In this case, Tobit does not apply; however, there is a similar maximum likelihood procedure, called **truncated regression**, for such a case. An advanced reference is William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, Section 19.2.3.

[27] The asymptotic covariance matrix can be found in *Introduction to the Theory and Practice of Econometrics, 2nd edition*, by George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee (John Wiley and Sons, 1988), Section 19.3.2.

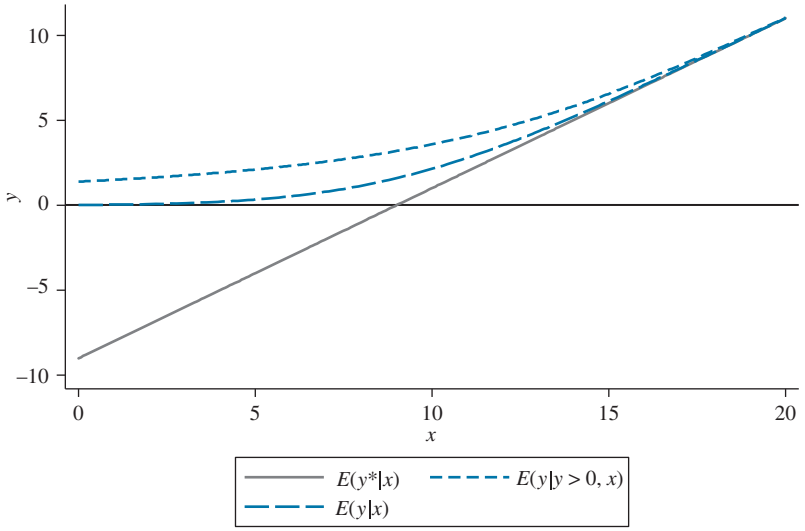


FIGURE 16.3 **Three regression functions.**

model, as is shown in Figure 16.3. The marginal effect can be decomposed into two factors called the "McDonald–Moffitt" decomposition:

$$\frac{\partial E(y|x)}{\partial x} = \text{Prob}(y > 0)\frac{\partial E(y|x, y > 0)}{\partial x} + E(y|x, y > 0)\frac{\partial \text{Prob}(y > 0)}{\partial x}$$

The first factor accounts for the marginal effect of a change in x for the portion of the population whose y-data is observed already. The second factor accounts for changes in the proportion of the population who switch from the y-unobserved category to the y-observed category when x changes.[28]

EXAMPLE 16.16 | Tobit Model of Hours Worked

An example that illustrates the situation is based on Thomas Mroz's (1987) study of married women's labor force participation and wages. The data are in the file *mroz* and consist of 753 observations on married women. Of these, 325 did not work outside the home and thus had no hours worked and no reported wages. The histogram of hours worked is shown in Figure 16.4. The histogram shows the large fraction of women with zero hours of work.

If we wish to estimate a model explaining the market hours worked by a married woman, what explanatory variables would we include? Factors that would tend to pull a woman into the labor force are her education and her prior labor market experience. Factors that may reduce her incentive to work are her age and the presence of young children in the home.[29]

Thus, we might propose the regression model

$$HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 AGE + \beta_4 KIDSL6 + e \quad (16.36)$$

[28] J. F. McDonald and R. A. Moffitt (1980) "The Uses of Tobit Analysis," *Review of Economics and Statistics*, 62, 318–321. Jeffrey M. Wooldridge (2009) *Introductory Econometrics: A Modern Approach, 5th edition*, South-Western Cengage Learning, Section 17.2 has a relatively friendly presentation.

[29] This equation does not include wages, which is jointly determined with hours. The model in (16.36) may be considered a reduced-form equation. See Section 11.2.

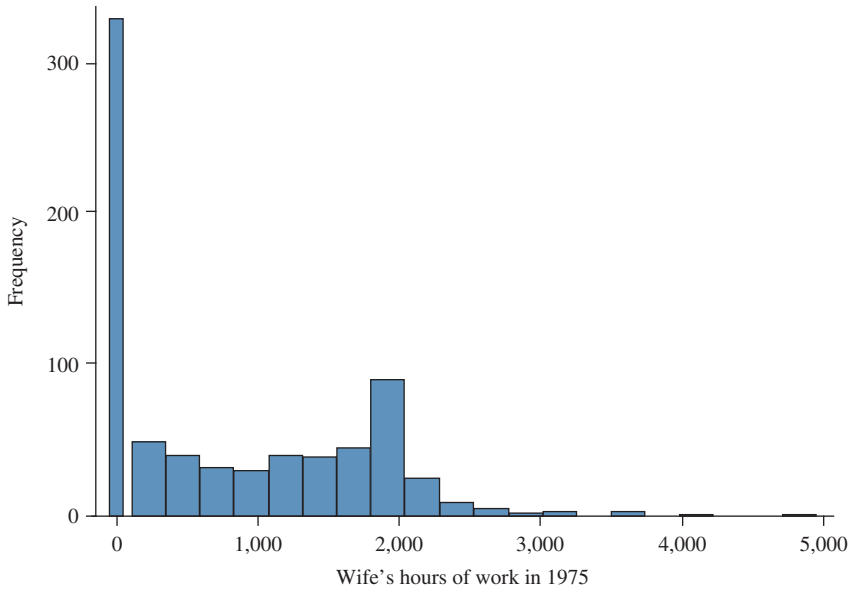


FIGURE 16.4 **Wife's hours of work in 1975.**

where *KIDSL6* is the number of children less than 6 years old in the household. Using Mroz's data, we obtain the estimates shown in Table 16.7. As previously argued, the least squares estimates are unreliable because the least squares estimator is both biased and inconsistent. The Tobit estimates have the anticipated signs and are all statistically significant at the 0.01 level. To compute the scale factor required for calculation of the marginal effects, we must choose values of the explanatory variables. We choose the sample means for *EDUC* (12.29), *EXPER* (10.63), and *AGE* (42.54) and assume one small child at home (rather than the mean value of 0.24). The calculated scale factor is $\tilde{\Phi} = 0.3630$. Thus, the marginal effect on observed hours of work of another year of education is

$$\frac{\partial E(HOURS)}{\partial EDUC} = \tilde{\beta}_2 \tilde{\Phi} = 73.29 \times 0.3630 = 26.61$$

That is, we estimate that another year of education will increase a wife's hours of work by about 27 hours, conditional upon the assumed values of the explanatory variables.

TABLE 16.7 **Estimates of Labor Supply Function**

Estimator	Variable	Estimate	Standard Error
Least squares	*INTERCEPT*	1335.31	235.65
	EDUC	27.09	12.24
	EXPER	48.04	3.64
	AGE	−31.31	3.96
	KIDSL6	−447.85	58.41
Least squares $y > 0$	*INTERCEPT*	1829.75	292.54
	EDUC	−16.46	15.58
	EXPER	33.94	5.01
	AGE	−17.11	5.46
	KIDSL6	−305.31	96.45
Tobit	*INTERCEPT*	1349.88	386.30
	EDUC	73.29	20.47
	EXPER	80.54	6.29
	AGE	−60.77	6.89
	KIDSL6	−918.92	111.66
	SIGMA	1133.70	42.06

16.7.5 Sample Selection

If you consult an econometrician concerning an estimation problem, the first question you will usually hear is, "How were the data obtained?" If the data are obtained by random sampling, then classic regression methods, such as least squares, work well. However, if the data are obtained by a sampling procedure that is not random, then standard procedures do not work well. Economists regularly face such data problems. A famous illustration comes from labor economics. If we wish to study the determinants of the wages of married women, we face a *sample selection* problem. If we collect data on married women and ask them what wage rate they earn, many will respond that the question is not relevant since they are homemakers. We only observe data on market wages when the woman chooses to enter the workforce. One strategy is to ignore the women who are not in the labor force, omit them from the sample, then use least squares to estimate a wage equation for those who work. This strategy fails, the reason for the failure being that our sample is not a random sample. The data we observe are "selected" by a systematic process for which we do not account.

A solution to this problem is a technique called **Heckit**, named after its developer, Nobel Prize winning econometrician James Heckman. This simple procedure uses two estimation steps. In the context of the problem of estimating the wage equation for married women, a probit model is first estimated explaining why a woman is in the labor force or not. In the second stage, a least squares regression is estimated relating the wage of a working woman to education, experience, and so on, and a variable called the "inverse Mills ratio," or IMR. The IMR is created from the first step probit estimation and accounts for the fact that the observed sample of working women is not random.

The econometric model describing the situation is composed of two equations. The first is the *selection equation* that determines whether the variable of interest is observed. The sample consists of N observations; however, the variable of interest is observed only for $n < N$ of these. The selection equation is expressed in terms of a latent variable z_i^* that depends on one or more explanatory variables w_i and is given by

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i, \quad i = 1, \ldots, N \tag{16.37}$$

For simplicity, we will include only one explanatory variable in the selection equation. The latent variable is not observed, but we do observe the indicator variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16.38}$$

The second equation is the linear model of interest. It is

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, n, \quad N > n \tag{16.39}$$

A **selectivity problem** arises when y_i is observed only when $z_i = 1$ and if the errors of the two equations are correlated. In such a situation, the usual least squares estimators of β_1 and β_2 are biased and inconsistent.

Consistent estimators are based on the conditional regression function[30]

$$E\left(y_i | z_i^* > 0\right) = \beta_1 + \beta_2 x_i + \beta_\lambda \lambda_i, \quad i = 1, \ldots, n \tag{16.40}$$

where the additional variable λ_i is the inverse Mills ratio. It is equal to

$$\lambda_i = \frac{\phi\left(\gamma_1 + \gamma_2 w_i\right)}{\Phi\left(\gamma_1 + \gamma_2 w_i\right)} \tag{16.41}$$

[30] Our Appendix B.2.6 provides a brief introduction to this important concept. See William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, Chapter 19.2 for much more about the truncated normal.

While the value of λ_i is not known, the parameters γ_1 and γ_2 can be estimated using a probit model, based on the observed binary outcome z_i in (16.38). Then the estimated IMR

$$\tilde{\lambda}_i = \frac{\phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}{\Phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}$$

is inserted into the regression equation as an extra explanatory variable, yielding the estimating equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_\lambda \tilde{\lambda}_i + v_i, \quad i = 1, \ldots, n \tag{16.42}$$

Least squares estimation of this equation yields consistent estimators of β_1 and β_2. A word of caution, however, as the least squares estimator is inefficient relative to the maximum likelihood estimator, and the usual standard errors and t-statistics produced after estimation of (16.42) are incorrect. Proper estimation of standard errors requires the use of specialized software for the "Heckit" model.

EXAMPLE 16.17 | Heckit Model of Wages

As an example, we will reconsider the analysis of wages earned by married women using the Mroz (1987) data in the data file *mroz*. In the sample of 753 married women, 428 have market employment and nonzero earnings. First, let us estimate a simple wage equation, explaining ln(*WAGE*) as a function of the woman's education, *EDUC*, and years of market work experience (*EXPER*), using the 428 women who have positive wages. The result is

$$\begin{aligned} \ln(WAGE) &= -0.4002 + 0.1095 EDUC \\ (t) \quad & (-2.10) \quad (7.73) \\ &+ 0.0157 EXPER \qquad R^2 = 0.1484 \qquad (16.43) \\ & (3.90) \end{aligned}$$

The estimated return to education is about 11%, and the estimated coefficients of both education and experience are statistically significant.

The Heckit procedure starts by estimating a probit model of labor force participation. As explanatory variables we use the woman's age, her years of education, an indicator variable for whether she has children, and the marginal tax rate that she would pay upon earnings if employed. The estimated probit model is

$$\begin{aligned} \widehat{P(LFP=1)} &= \Phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC \\ (t) \qquad & \qquad (-2.93) \qquad (3.61) \\ &- 0.3139 KIDS - 1.3939 MTR\big) \\ & (-2.54) \qquad (-2.26) \end{aligned}$$

As expected, the effects of age, the presence of children, and the prospects of higher taxes significantly reduce the probability that a woman will join the labor force, while education increases it. Using the estimated coefficients, we compute the inverse Mills ratio for the 428 women with market wages

$$\tilde{\lambda} = IMR = \frac{\phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC - 0.3139 KIDS - 1.3939 MTR\big)}{\Phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC - 0.3139 KIDS - 1.3939 MTR\big)}$$

This is then included in the wage equation, and least squares estimation applied to obtain

$$\begin{aligned} \ln(WAGE) &= 0.8105 + 0.0585 EDUC + 0.0163 EXPER \\ (t) \quad & (1.64) \quad (2.45) \qquad (4.08) \\ (t-adj) \quad & (1.33) \quad (1.97) \qquad (3.88) \\ &- 0.8664 IMR \\ & (-2.65) \\ & (-2.17) \qquad (16.44) \end{aligned}$$

Two results are of note. First, the estimated coefficient of the inverse Mills ratio is statistically significant, implying that there is a **selection bias** present in the least squares results (16.43). Second, the estimated return to education has fallen from approximately 11% to approximately 6%. The upper row of t-statistics is based on standard errors as usually computed when using least squares regression. The usual standard errors do not account for the fact that the inverse Mills ratio is itself an estimated value. The correct standard errors,[31] which do account for the first stage probit

[31] The formulas are very complicated. See William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, p. 954. There are several software packages, such as Stata and LIMDEP, that report correct standard errors.

estimation, are used to construct the "adjusted t-statistics" reported in (16.44). As you can see the adjusted t-statistics are slightly smaller, indicating that the adjusted standard errors are somewhat larger than the usual ones.

In most instances, it is preferable to estimate the full model, both the selection equation and the equation of interest, jointly by maximum likelihood. While the nature of this procedure is beyond the scope of this book, it is available in some software packages. The maximum likelihood estimated wage equation is

$$\underset{(t)}{\ln(WAGE)} = \underset{(2.84)}{0.6686} + \underset{(3.96)}{0.0658}EDUC + \underset{(2.87)}{0.0118}EXPER$$

The standard errors based on the full information maximum likelihood procedure are smaller than those yielded by the two-step estimation method.

16.8 Exercises

16.8.1 Problems

16.1 In Examples 16.2 and 16.4, we presented the linear probability and probit model estimates using an example of transportation choice. The logit model for the same example is $P(AUTO = 1) = \Lambda(\gamma_1 + \gamma_2 DTIME)$, where $\Lambda(\bullet)$ is the logistic *cdf* in equation (16.7). The logit model parameter estimates and their standard errors are

$$\underset{(se)}{\tilde{\gamma}_1 + \tilde{\gamma}_2 DTIME} = \underset{(0.7505)}{-0.2376} + \underset{(0.2064)}{0.5311}DTIME$$

a. Calculate the estimated probability that a person will choose automobile transportation given that $DTIME = 1$.

b. Using the probit model results in Example 16.4, calculate the estimated probability that a person will choose automobile transportation given that $DTIME = 1$. How does this result compare to the logit estimate? [*Hint*: Recall that Statistical Table 1 gives cumulative probabilities for the standard normal distribution.]

c. Using the logit model results, compute the estimated marginal effect of an increase in travel time of 10 minutes for an individual whose travel time is currently 30 minutes longer by bus (public transportation). Using the linear probability model results, compute the same marginal effect estimate. How do they compare?

d. Using the logit model results, compute the estimated marginal effect of a decrease in travel time of 10 minutes for an individual whose travel time is currently 50 minutes longer by driving. Using the probit results, compute the same marginal effect estimate. How do they compare?

16.2 In Appendix 16A.1, we illustrate the calculation of a standard error for the marginal effect in a probit model of transportation, Example 16.4. In the appendix, the calculation is for the marginal effect when it currently takes 20 minutes longer to commute by bus ($DTIME = 2$).

a. Repeat the calculation for the probit model when $DTIME = 1$. [*Hint*: The values of the standard normal *pdf* are given in Statistical Table 6.]

b. Using the probit model, construct a 95% interval estimate for the marginal effect of a 10-minute increase in travel time by bus when $DTIME = 1$.

c. The logit model estimates and standard errors are

$$\underset{(se)}{\tilde{\gamma}_1 + \tilde{\gamma}_2 DTIME} = \underset{(0.7505)}{-0.2376} + \underset{(0.2064)}{0.5311}DTIME$$

The estimated coefficient covariance is $\widehat{\text{cov}}(\tilde{\gamma}_1, \tilde{\gamma}_2) = -0.025498$. Calculate the standard error of the marginal effect of a 10-minute increase in travel time when $DTIME = 1$. [*Hint*: Carry through the steps in Appendix 16A.1 using equation (16.17) in place of $\Phi(\cdot)$ and equation (16.16) in place of $\phi(\cdot)$.]

d. Construct a 95% interval estimate for the marginal effect of a 10-minute increase in travel time by bus, when $DTIME = 1$ for the logit model.

16.3 In Example 16.3, we illustrate the calculation of the likelihood function for the probit model in a small example.

- **a.** Calculate the probability that $y = 1$ if $x = 1.5$, given the values of the maximum likelihood estimates.
- **b.** Using the threshold 0.5 and the result in part (a), predict the value of y if $x = 1.5$, the first observation, given the values of the maximum likelihood estimates. Does your prediction agree with the actual outcome $y = 1$?
- **c.** Calculate the value of the likelihood function, illustrated in equation (16.14), using the given $N = 3$ data pairs, if the parameter values are $\beta_1 = -1$ and $\beta_2 = 0.2$. Compare this value to the value of the likelihood function evaluated at the maximum likelihood estimates, given in Example 16.3. Which is larger?
- **d.** For the probit model, the value of the likelihood function (16.14) will always be between zero and one. True or false? Explain.
- **e.** For the probit model, the value of the log-likelihood function (16.15) will always be negative. True or false? Explain.

16.4 In Example 16.3, we illustrate the calculation of the likelihood function for the probit model in a small example. In this exercise, we will repeat that example using logit instead of probit. The logit model for the same example is $P(y = 1) = \Lambda(\gamma_1 + \gamma_2 x)$, where $\Lambda(\bullet)$ is the logistic *cdf* in equation (16.7). The maximum likelihood estimates of the parameters are $\tilde{\gamma}_1 + \tilde{\gamma}_2 x = -1.836 + 3.021x$. The maximized value of the log-likelihood function is -1.612.

- **a.** Calculate the probability that $y = 1$ if $x = 1.5$, given the values of the maximum likelihood estimates.
- **b.** Using the threshold 0.5 and the result in part (a), predict the value of y if $x = 1.5$, the first observation, given the values of the maximum likelihood estimates. Compare your prediction to the actual outcome $y = 1$ in the first observation.
- **c.** Calculate the value of the likelihood function, illustrated in equation (16.14) but substituting equation (16.17) in place of $\Phi(\bullet)$ and using the given $N = 3$ data pairs, if the parameter values are $\gamma_1 = -1$ and $\gamma_2 = 2$. Compare this value to the value of the likelihood function evaluated at the maximum likelihood estimates. Which is larger?
- **d.** For the logit model, the value of the likelihood function (16.14), with $\Lambda(\bullet)$ in place of $\Phi(\bullet)$, will always be between zero and one. True or false? Explain.
- **e.** For the logit model, the value of the log-likelihood function (16.15), with $\Lambda(\bullet)$ in place of $\Phi(\bullet)$, will always be negative. True or false? Explain.

16.5 We are given three observations on binary choice with $y_1 = 1, y_2 = 1, y_3 = 0$. Consider a logit model with only an intercept, $P(y = 1) = \Lambda(\gamma_1)$, where $\Lambda(\bullet)$ is the logistic *cdf*.

- **a.** Show that the log-likelihood function is $\ln L(\gamma_1) = 2\ln\Lambda(\gamma_1) + \ln[1 - \Lambda(\gamma_1)]$.
- **b.** Show that $d\ln L(\gamma_1)/d\gamma_1 = 2\lambda(\gamma_1)/\Lambda(\gamma_1) - \lambda(\gamma_1)/[1 - \Lambda(\gamma_1)]$, where $\lambda(\cdot)$ is the logistic *pdf* in (16.14). [*Hint*: Use Derivative Rules 8 and 9 from Appendix A.3.]
- **c.** The value of γ_1 such that $d\ln L(\gamma_1)/d\gamma_1 = 0$ is the maximum likelihood estimator $\tilde{\gamma}_1$. True, false, or maybe?
- **d.** It can be shown that for the logit model $\ln L(\gamma_1)$ is strictly concave, meaning that the second derivative is negative for all values of γ_1 or $d^2\ln L(\gamma_1)/d\gamma_1{}^2 < 0$. What is your answer to (c) now? [*Hint*: See Appendix A.3.4.]
- **e.** Setting the derivative in (c) to zero and solving, show that $\Lambda(\tilde{\gamma}_1) = 2/3$. [*Note*: This does not require you to first solve for $\tilde{\gamma}_1$.]
- **f.** Now, solve the condition in (c) to show that $\tilde{\gamma}_1 = -\ln(1/2)$.

16.6 In this exercise, we generalize the results in Exercise 16.5. Consider a logit model with only an intercept, $P(y = 1) = \Lambda(\gamma_1)$, where $\Lambda(\bullet)$ is the logistic *cdf*. Suppose in a sample of N observations, there are N_1 values $y_i = 1$ and N_0 values $y_i = 0$.

- **a.** Show that the logit log-likelihood function is $\ln L(\gamma_1) = N_1\ln\Lambda(\gamma_1) + N_0\ln[1 - \Lambda(\gamma_1)]$.
- **b.** Show that $d\ln L(\gamma_1)/d\gamma_1 = N_1\lambda(\gamma_1)/\Lambda(\gamma_1) - N_0\lambda(\gamma_1)/[1 - \Lambda(\gamma_1)]$, where $\lambda(\cdot)$ is given in (16.6). [*Hint*: Use Derivative Rules 8 and 9 from Appendix A.3.]
- **c.** Setting the derivative in (b) to zero and solving, show that $\Lambda(\tilde{\gamma}_1) = N_1/N$. What is the interpretation of N_1/N? [*Note*: This does not require you to first solve for $\tilde{\gamma}_1$, the MLE.]

d. Using (c) show that $\ln L(\tilde{\gamma}_1) = N_1 \ln(N_1/N) + N_0 \ln(N_0/N)$.

e. Show that a probit model, $P(y = 1) = \Phi(\gamma_1)$, where $\Phi(\bullet)$ is the standard normal *cdf*, results in the same value for the log-likelihood as in part (d).

16.7 Exercise 16.5 shows that given the three observations on binary choice with $y_1 = 1, y_2 = 1, y_3 = 0$ the maximum likelihood estimator of the logit model $P(y = 1) = \Lambda(\gamma_1)$ is $\tilde{\gamma}_1 = -\ln(1/2) = 0.6931472$ and that $\Lambda(\tilde{\gamma}_1) = 2/3$.

a. Using these results show that $\ln L(\tilde{\gamma}_1) = 2\ln\Lambda(\tilde{\gamma}_1) + \ln[1 - \Lambda(\tilde{\gamma}_1)] = -1.9095425$.

b. Using the data in Example 16.3, and the logit model $P(y = 1|x) = \Lambda(\gamma_1 + \gamma_2 x)$, we find that the maximum likelihood estimates of the parameters are $\tilde{\gamma}_1 + \tilde{\gamma}_2 x = -1.836 + 3.021x$, and the maximized value of the log-likelihood function is -1.612. Using these results, and those in (a), carry out the likelihood ratio test of H_0: $\gamma_2 = 0$ versus H_1: $\gamma_2 \neq 0$ at the 5% level of significance.

c. Calculate the *p*-value for the test in (b).

16.8 Consider a probit model designed to explain the choice by homebuyers of fixed versus adjustable rate mortgages. The explanatory variables, with sample means in parentheses, are *FIXRATE* (13.25) = fixed interest rate; *MARGIN* (2.3) = the variable rate − the fixed rate; and *NETWORTH* (3.5) = borrower's net worth ($100,000 units). The dependent variable is *ADJUST* (0.41) = 1 if an adjustable mortgage is chosen. The coefficient estimates, in Table 16.8, use 78 observations over the period January, 1983 to February, 1984.

TABLE 16.8 **Estimates for Exercise 16.8**

	C	*FIXRATE*	*MARGIN*	*NETWORTH*	ln*L*(Model)	ln*L*(*C*)
Model 1	−7.0166	0.5301	−0.2675	0.0864	−42.0625	−52.8022
(se)	(3.3922)	(0.2531)	(0.1304)	(0.0354)		
Model 2	−9.8200	0.7535	−0.1945		−45.1370	−52.8022
(se)	(3.1468)	(0.2328)	(0.1249)			

a. What information is provided by the signs of the estimated coefficients of Model 1? Are the signs consistent with economic reasoning? Which coefficients are significant at the 5% level?

b. Carry out a likelihood ratio test of the model significance at the 1% level for Model 1. In Table 16.8, ln*L*(Model) is the log-likelihood of the full model and ln*L*(*C*) is the log-likelihood of the model including only the constant term.

c. What is the estimated probability of a borrower choosing an adjustable rate mortgage if *FIXRATE* = 12, *MARGIN* = 2, and *NETWORTH* = 3? What is the estimated probability of a borrower choosing an adjustable rate mortgage if *FIXRATE* = 12, *MARGIN* = 2, and *NETWORTH* = 10?

d. Carry out a likelihood ratio test of the hypothesis that *NETWORTH* has no effect on the choice of mortgage type, against the alternative that it does, at the 1% level.

e. Using Model 2, what is the marginal effect of *MARGIN* on the probability of choosing an adjustable rate mortgage if *FIXRATE* = 12 and *MARGIN* = 2?

f. Using Model 2, calculate the discrete change in the probability of choosing an adjustable rate mortgage if *MARGIN* increases from 2% to 4%, while *FIXRATE* remains 12%? Is the value twice the value found in part (e)?

16.9 Consider a probit model explaining the choice to attend college by high-school graduates. Define *COLLEGE* = 1 if a high-school graduate chooses either a 2-year or 4-year college, and zero otherwise. We use explanatory variables *GRADES*, 13 point scale with 1 indicating highest grade (A+) and 13 the lowest (F); *FAMINC*, gross family income in $1000 units; and *BLACK* = 1 if black. Using a sample of $N = 1000$ graduates the estimated model is

$$\underset{}{P(COLLEGE = 1)} = \underset{}{\Phi}\underset{}{(2.5757} \underset{(0.0265)}{- 0.3068GRADES} + \underset{(0.0017)}{0.0074FAMINC} + \underset{(0.2177)}{0.6416BLACK}\underset{}{)}$$
$$\text{(se)}$$

a. What information is provided by the signs of the estimated coefficients? Which coefficients are statistically significant at the 5% level?

b. Estimate the probability of attending college for a white student with *GRADES* = 2 (A) and *FAMINC* = 50 ($50,000). Repeat this probability calculation if *GRADES* = 5 (B).

c. Estimate the probability of attending college for a black student with *GRADES* = 5 (B) and *FAMINC* = 50 ($50,000). Compare this probability to the comparable probability for a white student calculated in part (b).

d. Calculate the marginal effect of an increase in family income of $1000 on the probability of attending college for a white student with *GRADES* = 5 (B).

e. The log-likelihood for the model estimated above is −423.36. Omitting *FAMINC* and *BLACK* the log-likelihood of the estimated probit model is −438.26. Test the joint significance of *FAMINC* and *BLACK* at the 1% level of significance using a likelihood ratio test.

16.10 Consider a probit model explaining the choice to attend a 4-year college rather than a 2-year college by high-school graduates who chose to attend a postsecondary school. Define *FOURYR* = 1 if a high-school graduate chooses 4-year college and *FOURYR* = 0 if the high school graduate chooses a 2-year college. We use explanatory variables *GRADES*, 13 point scale with 1 indicating highest grade (A+) and 13 the lowest (F); *FAMINC*, gross family income in $1000 units; and *HSCATH* = 1 if the student attended a Catholic high school and *HSCATH* = 0 otherwise. Table 16.9 contains some probit model estimates.

TABLE 16.9 Estimates for Exercise 16.10

Model	(1)	(2)	(3)	(4)	(5)
				HSCATH = 0	*HSCATH* = 1
C	1.6395	1.6299	1.6039	1.6039	2.3143
	(23.8658)	(23.6925)	(22.5893)	(22.5893)	(8.0379)
GRADES	−0.2350	−0.2357	−0.2344	−0.2344	−0.2603
	(−25.1058)	(−25.1437)	(−24.2364)	(−24.2364)	(−6.7691)
FAMINC	0.0042	0.0040	0.0043	0.0043	0.0015
	(8.2798)	(7.6633)	(7.7604)	(7.7604)	(1.0620)
HSCATH		0.3645	0.7104		
		(5.0842)	(2.3954)		
HSCATH × *GRADES*			−0.0259		
			(−0.6528)		
HSCATH × *FAMINC*			−0.0028		
			(−1.9050)		
N	5254	5254	5254	4784	470
ln*L*	−2967.91	−2954.50	−2952.68	−2735.14	−217.54

t-statistics in parentheses.

a. Using Model (2), how large an effect on the probability of attending a 4-year college does attending a catholic high school have for a student with *GRADES* = 5 (B) and family income of $100,000.

b. Comparing Models (2) and (3), are the interaction variables *HSCATH* × *GRADES* and *HSCATH* × *FAMINC* jointly significant at 5% using a likelihood ratio test?

c. Can we interpret the Model (3) results as saying an increase in family income reduces the probability of attending a 4-year college for someone graduating from a Catholic high school? What is the marginal effect of an additional $1000 in family income for a Catholic high school student with *GRADES* = 5 (B) and family income of $50,000?

d. Using Model (3), compute the probability of attending a 4-year college for someone graduating from a Catholic high school with *GRADES* = 5 (B) and family income of $100,000. Compare this probability to a student who did not attend a Catholic high school but has *GRADES* = 5 (B) and family income of $100,000.

e. Using Models (1) and (3), test the null hypothesis that the probit model parameters are the same for students who attend and do not attend a Catholic high school. Use a likelihood ratio test at the 5% level of significance.

f. Using Models (4) and (5), estimate the probit model separately for *HSCATH* = 0 and *HSCATH* = 1. Compute the sum of the log-likelihood functions values. Compare the sum to the log-likelihood for Model (3). Algebraically show that this is not an accident.

16.11 Using data on $N = 4{,}642$ infant births, we estimate a probit model with dependent variable *LBWEIGHT* = 1 if it is a low birthweight baby and 0 otherwise, *MAGE* is the mother's age, *PRENATAL1* = 1 if first prenatal visit is in 1 trimester and 0 otherwise, and *MBSMOKE* = 1 if the mother smoked and 0 otherwise. The results are in Table 16.10.

TABLE 16.10 Probit Estimates for Exercise 16.11

	C	*MAGE*	*PRENATAL1*	*MBSMOKE*	$MAGE^2$
Model 1	−1.2581	−0.0103	−0.1568	0.3974	
(se)	(0.1436)	(0.0054)	(0.0710)	(0.0670)	
Model 2	−0.1209	−0.1012	−0.1387	0.4061	0.0017
(se)	(0.4972)	(0.0385)	(0.0716)	(0.0672)	(0.0007)

a. In Model 1, comment on estimated signs and significance of the coefficients on *PRENATAL1* and *MBSMOKE*.

b. Using Model 1, calculate the marginal effect on the probability of a low birthweight baby given an increase in the mother's age by 1 year, for a woman who is 20 years old with *PRENATAL1* = 0 and *MBSMOKE* = 0. Repeat this calculation for a woman who is 50 years old. Do the results make sense?

c. Using Model 2, calculate the marginal effect on the probability of a low birthweight baby given an increase in the mother's age by 1 year, for a woman who is 20 years old with *PRENATAL1* = 0 and *MBSMOKE* = 0. Repeat this calculation for a woman who is 50 years old. Compare these results to those in part (b).

d. Using Model 2, calculate the impact of a prenatal visit in the first trimester on the probability of having a low birthweight baby for a woman who is 30 years old and smokes.

e. Using Model 2, calculate the impact of a mother smoking on the probability having a low birthweight baby given that she is 30 years old and had a prenatal visit in the first trimester.

f. Using Model 2, calculate the age at which the probability of a low birthweight baby is a minimum.

16.12 This exercise is an extension of Example 16.12 using the larger data set *nels* with 6,649 observations. Two estimated multinomial logit models are reported in Table 16.11. In addition to the variable *GRADES*, we have *FAMINC* = family income ($1000 units) and indicator variables for sex and race. The baseline group is students who chose not to attend college.

TABLE 16.11 Estimates for Exercise 16.12

	Model 1		Model 2	
PSECHOICE	Coefficient	*t*-value	Coefficient	*t*-value
2				
C	1.7101	9.3293	1.9105	11.1727
GRADES	−0.2711	−13.1969	−0.2780	−13.9955
FAMINC	0.0124	8.3072	0.0116	8.0085
FEMALE	0.2284	3.0387		
BLACK	0.0554	0.4322		
3				
C	4.6008	25.7958	4.6111	27.8351
GRADES	−0.6895	−32.2723	−0.6628	−32.3721
FAMINC	0.0200	13.5695	0.0183	12.9450
FEMALE	0.0422	0.5594		
BLACK	0.9924	8.0766		
ln(*L*)	−5699.8023		−5751.5982	

a. Which estimated coefficients are significant in Model 1? Based on the *t*-values, should we consider dropping *FEMALE* and *BLACK* from the model?

b. Compare the results of Model 1 to Model 2 using a likelihood ratio test. Using the $\alpha = 0.01$ level of significance, can we reject the null hypothesis that the Model 1 coefficients of *FEMALE* and *BLACK* are zero?

c. Compute the estimated probability that a white male student with *GRADES* = 5 (B) and *FAMINC* of $100,000 will attend a 4-year college.

d. Compute the odds, or probability ratio, that a white male student with *GRADES* = 5 (B) and *FAMINC* of $100,000 will attend a 4-year college rather than not attend any college.

e. Compute the change in probability of attending a 4-year college for a white male student with median *FAMINC* =$100,000 whose *GRADES* change from 5 (B) to 2 (A).

16.13 This exercise is an extension of Example 16.13. It is a conditional logit model of choice among 3 brands of soda: Coke, Pepsi, and 7-Up. The data are in the data file *cola*. As in the example, we choose Coke to be the base alternative, setting its alternative specific constant (intercept) to zero. We add to the model indicator variables *FEATURE*, indicating whether the product was "featured" at the time, and *DISPLAY* for whether there was a store display at the time of purchase. The model estimates are in Table 16.12.

TABLE 16.12 Estimates for Exercise 16.13

	Model 1		Model 2	
	Coefficient	*t*-Statistic	Coefficient	*t*-Statistic
PRICE	−1.7445	−9.6951	−1.8492	−9.8017
FEATURE	−0.0106	−0.1327	−0.0409	−0.4918
DISPLAY	0.4624	4.9700	0.4727	5.0530
PEPSI			0.2841	4.5411
7-UP			0.0907	1.4173
ln(*L*)	−1822.2267		−1811.3543	

a. Using Model 1, calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.25, Coke has a display but Pepsi does not, and neither are featured. Note that the model contains no alternative specific constants.

b. Using Model 1, calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.00, Coke has a display but Pepsi does not, and neither are featured.

c. Compute the change in the probability of purchase of each type of soda if the price of Coke changes from \$1.25 to \$1.50, with the prices of Pepsi and 7-Up remaining at \$1.25. Assume that a display is present for Coke, but not for the others, and none of the items is featured.

d. In Model 2, we add the alternative specific "intercept" terms for Pepsi and 7-Up to the Model 1. Calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.25, Coke has a display but Pepsi does not, and neither are featured.

e. Using Model 2, compute the change in the probability of purchase of each type of soda if the price of Coke changes from \$1.25 to \$1.50, with the prices of Pepsi and 7-Up remaining at \$1.25. Assume that a display is present for Coke, but not for the others, and none of the items is featured.

f. The value of the log-likelihood function for the model in Example 16.13 is −1824.5621. Test the null hypothesis that the coefficients of the marketing variables, *FEATURE* and *DISPLAY*, are zero, against the alternative that they are not, using a likelihood ratio test with $\alpha = 0.01$.

16.14 In Example 16.14, we described an ordinal probit model for postsecondary education choice, and estimated a simple model in which the choice depended simply on the student's *GRADES*. Expand the ordered probit model to include family income (*FAMINC*, in \$1000), family size (*FAMSIZ*), the dummy variables *BLACK* and *PARCOLL* = 1 if a parent has at least a college degree, and 0 otherwise. The estimates of this model are Model 2 in Table 16.13.

TABLE 16.13 **Estimates for Exercise 16.14**

	Model 1		Model 2	
PSECHOICE	Coefficient	Standard Error	Coefficient	Standard Error
GRADES	−0.3066	0.0192	−0.2953	0.0202
FAMINC			0.0053	0.0013
FAMSIZ			−0.0241	0.0302
BLACK			0.7131	0.1768
PARCOLL			0.4236	0.1016
$\hat{\mu}_1$	−2.9456	0.1468	−2.5958	0.2046
$\hat{\mu}_2$	−2.0900	0.1358	−1.6946	0.1971
ln*L*	−875.8217		−839.8647	

a. Using the estimates in Table 16.13, Model 1, calculate the probability that a student will choose no college, a 2-year college, and a 4-year college if the student's grades are *GRADES* = 7 (B−). Recompute these probabilities assuming that *GRADES* = 3 (A−). Discuss the probability changes. Are they what you anticipated? Explain.

b. Discuss the Model 2 estimates, their signs and significance. [*Hint*: recall that the sign indicates the direction of the effect for the highest category but is opposite for the lowest category].

c. Test the joint significance of the variables added in (b) using a likelihood ratio test at the 1% level of significance.

d. Compute the probability that a black student from a household of four members with \$100,000 income, and with at least one parent having at least a college degree, so that *PARCOLL* = 1, will attend a 4-year college if (i) *GRADES* = 7 and (ii) *GRADES* = 3.

e. Repeat (d) for a "nonblack" student and discuss the differences in your findings.

16.15 Consider a Poisson regression explaining the number of Olympic Games medals won using data from 1988 (in Seoul, South Korea) and 1992 (in Barcelona, Spain) by various

countries as a function of $LPOP = \ln(POP)$ = the logarithm of population in millions, and $LGDP = \ln(GDP)$ = the logarithm of gross domestic product (in billions of 1995 dollars). That is, $E(MEDALTOT|\mathbf{X}) = \exp[\beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP)]$. The estimated coefficients, using 316 observations, are in Table 16.14, Model 1.

TABLE 16.14 **Estimates for Exercise 16.15**

	Model 1		Model 2	
	Coefficient	Standard Error	Coefficient	Standard Error
C	−1.4442	0.0826	−1.4664	0.0835
LPOP	0.2143	0.0217	0.2185	0.0219
LGDP	0.5556	0.0164	0.5536	0.0165
HOST			0.6620	0.1375

a. Using Model 1 results, what is the estimated impact on the number of medals won if *GDP* increases by 1%? [*Hint*: It can be shown (can you?) that β_3 is an elasticity.]

b. In 1996, Bulgaria had *GDP* = 11.8 billion and a population of 8.356 million. Estimate the expected number of medals that Bulgaria would win in the Olympics, held in Atlanta, USA. They did win 15 medals.

c. Calculate the probability that Bulgaria in 1996 would win one or fewer medals.

d. In 1996, Switzerland had *GDP* = 306 billion and a population of 6.875 million. Estimate the expected number of medals that Switzerland would win. They did win 1 medal.

e. Calculate the probability that Switzerland in 1996 would win one or fewer medals.

f. *HOST* is an indicator variable = 1 for the country hosting the Olympics. This variable is added in Model 2. Interpret its coefficient. [*Hint*: What is the estimated percentage change in the conditional mean?] Is the estimated effect large or small? Is the coefficient statistically significant at the 1% level?

g. In 1996, the Olympic games were held in the U.S. city of Atlanta, GA. In that year, the U.S. population was 265 million and its *GDP* was 7280 billion. Estimate the expected number of medals the United States would win using Model 1 and again using Model 2. The United States won 101 medals that year. Which model's estimated value was closer to the true outcome?

16.16 Consider a regression explaining the share of Olympic Games medals won by each country in 1988 (in Seoul, South Korea), 1992 (in Barcelona, Spain), and 1996 (in Atlanta, GA, USA) as a function of $LPOP = \ln(POP)$ = the logarithm of population in millions, $LGDP = \ln(GDP)$ = the logarithm of gross domestic product (in billions of 1995 dollars), and *HOST*, an indicator variable = 1 for the country hosting the Olympics. The total number of medals awarded in 1988 was 738; in 1992, there were 815 medals awarded, and in 1996, 842 medals were awarded. Using the total number of medals awarded, we compute the percentage share of medals (*SHARE*) won by each country.

a. The least squares estimates of $SHARE = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + \beta_4 HOST + e$ are in Table 16.15. Are the signs and significance of the coefficient estimates reasonable?

TABLE 16.15 **Estimates for Exercise 16.16**

	OLS			Tobit	
	Coefficient	Standard Error	HCE	Coefficient	Standard Error
C	−0.2929	0.1000	0.0789	−4.2547	0.3318
LPOP	−0.0058	0.0496	0.0352	0.1707	0.1135
LGDP	0.3656	0.0454	0.0579	0.9605	0.0973
HOST	4.1723	0.9281	2.0770	3.2475	1.4611
$\tilde{\sigma}$				2.4841	0.1273

b. Using the OLS estimates, what is the predicted effect of *GDP* on the expected share of medals won? That is, how much do we predict the share of medals won will change if *GDP* increases by 1%? Construct a 95% interval estimate of this effect.

c. For the model estimated by OLS, the robust Breusch-Pagan LM test statistic for heteroskedasticity as a function of ln(*GDP*) is $NR^2 = 32.80$. What can we conclude about the OLS estimator and the usual standard errors based on this test?

d. We also report the OLS heteroskedasticity robust standard errors (HCE) in Table 16.15. Construct a 95% interval estimate for the predicted effect of a 1% increase in *GDP* on the share of medals won using the robust standard errors.

e. Among the 508 countries competing in these summer Olympics, almost 62% won no medals. Does this cause any potential problems for the least squares estimator? By using robust standard errors in part (c), we have solved any problems with the OLS estimator. True or false? Explain your choice.

f. Compare the Tobit parameter estimates reported in Table 16.15 to the OLS estimates and standard errors. What are the differences? Is Tobit a reasonable estimator for the share of medals won in this example? Why?

g. Using the Tobit estimates, what is the estimated effect of *GDP* on the expected share of medals won for a nonhost country with $GDP = 150$ billion and $POP = 30$ million? That is, how much do we estimate the expected share of medals won will change if *GDP* increases by one percent? [*Hint*: In equation (16.35), let $y = SHARE$ and $x = \ln(GDP)$. Then

$$\partial E(SHARE|\mathbf{X})/\partial \ln(GDP) = \beta_3 \Phi\left[\frac{\beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + \beta_4 HOST}{\sigma}\right]$$

Also, $\partial \ln(GDP)/\partial GDP = 1/GDP$. Then refer to the analysis of the linear-log model in Section 4.3.3.]

16.8.2 Computer Exercises

16.17 In Chapter 7, we examined the Tennessee's Project STAR. In the experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular sized classes with 22–25 students, and regular sized classes with a full-time teacher aide to assist the teacher. In Example 7.11, we checked for random assignment of children to the three types of classes using a linear probability model, regressing the indicator *SMALL* (small class) on student characteristics. Let us reconsider this regression using logit rather than the linear probability model. If there is random assignment of children to types of classes, then we should not find any significant relationships. Use data file *star5_small2* for this exercise. The data file *star5* contains more observations.

a. Estimate a logit model with outcome variable *SMALL* and explanatory variables *BOY* and *BLACK*. Individually test the coefficients of these variables for significance. What do you find? Test the coefficients jointly for significance using the likelihood ratio test. What do you find? Can we reject the null hypothesis that assignment to small classes is done randomly?

b. Repeat the estimation and testing in part (a) using outcome variables *AIDE* and *REGULAR*. Do you find any evidence that students were not randomly assigned?

c. Add the variable *FREELUNCH* to the models in (a) and (b) and reestimate them. Do you find any evidence that there is a systematic pattern between class assignment and this variable?

d. Add the two variables *TCHWHITE* and *TCHMASTERS* to the models in (c) and reestimate them. In each, carry out a likelihood ratio test for the joint significance of *TCHWHITE* and *TCHMASTERS*. What do you conclude? In the experiment students were randomized within schools but not across schools. Does this offer any explanation of your findings? If so, how?

16.18 Mortgage lenders are interested in determining borrower and loan characteristics that may lead to delinquency or foreclosure. In the data file *lasvegas* are 1000 observations on mortgages for single family homes in Las Vegas, Nevada during 2008. The variable of interest is *DELINQUENT*, an indicator variable = 1 if the borrower missed at least three payments (90 + days late), but 0 otherwise. Explanatory variables are *LVR* = the ratio of the loan amount to the value of the property; *REF* = 1 if purpose of the loan was a "refinance" and = 0 if loan was for a purchase; *INSUR* = 1 if mortgage carries mortgage insurance, 0 otherwise; *RATE* = initial interest rate of the mortgage;

AMOUNT = dollar value of mortgage (in \$100,000); *CREDIT* = credit score, *TERM* = number of years between disbursement of the loan and the date it is expected to be fully repaid, *ARM* = 1 if mortgage has an adjustable rate, and = 0 if mortgage has a fixed rate.

a. Estimate the linear probability (regression) model explaining *DELINQUENT* as a function of the remaining variables. Use White heteroskedasticity robust standard errors. Are the signs of the estimated coefficients reasonable?

b. Use logit to estimate the model in (a). Are the signs and significance of the estimated coefficients the same as for the linear probability model?

c. Compute the predicted value of *DELINQENT* for the 500th and 1000th observations using both the linear probability model and the logit model. Interpret the values.

d. Construct a histogram of *CREDIT*. Using both linear probability and logit models, calculate the probability of delinquency for *CREDIT* = 500, 600, and 700 for a loan of \$250,000 (*AMOUNT* = 2.5). For the other variables, let the loan to value ratio (*LVR*) be 80%, the initial interest rate is 8%, all indicator variables take the value 0, and *TERM* = 30. Discuss similarities and differences among the predicted probabilities from the two models.

e. Using both linear probability and logit models, compute the marginal effect of *CREDIT* on the probability of delinquency for *CREDIT* = 500, 600, and 700, given that the other explanatory variables take the values in (d). Discuss the interpretation of the marginal effect.

f. Construct a histogram of *LVR*. Using both linear probability and logit models, calculate the probability of delinquency for *LVR* = 20 and *LVR* = 80, with *CREDIT* = 600 and other variables set as they are in (d). Compare and contrast the results.

g. Compare the percentage of correct predictions from the linear probability model and the logit model using a predicted probability of 0.5 as the threshold.

h. As a loan officer, you wish to provide loans to customers who repay on schedule and are not delinquent. Suppose you have available to you the first 500 observations in the data on which to base your loan decision on the second 500 applications (501–1,000). Is using the logit model with a threshold of 0.5 for the predicted probability the best decision rule for deciding on loan applications? If not, what is a better rule?

16.19 Mortgage lenders are interested in determining borrower and loan factors that may lead to delinquency or foreclosure. In the data file *vegas5_small* are 1000 observations on mortgages for single family homes in Las Vegas, Nevada during 2010. (The data file *vegas5* contains 10,000 observations.) The variable of interest is *DEFAULT*, an indicator variable = 1 if the borrower's payment was 90 + days late, but 0 otherwise. Explanatory variables are *ARM* = 1 if it's an adjustable rate mortgage, 0 if fixed; *REFINANCE* = 1 if loan is for a refinance of any type, 0 if for purchase; *LIEN2* = 1 if there is a second lien mortgage, 0 if it is the first lien; *TERM30* = 1 if it is a 30-year mortgage, 0 if 15-year mortgage; *UNDERWATER* = 1 if borrower estimated to owe more than the property is worth at time of origination, 0 otherwise; *LTV* = loan to value ratio of property at origination, percent; *RATE* = current interest rate on loan, percent; *AMOUNT* = loan amount in \$10,000 units; and *FICO* = borrower's credit score at origination.

a. Estimate the linear probability (regression) model explaining *DEFAULT* as a function of the remaining variables. Use White robust standard errors. Are the signs of the estimated coefficients reasonable?

b. Use probit to estimate the model in (a). Are the signs and significance of the estimated coefficients the same as for the linear probability model?

c. Compute the predicted value of *DEFAULT* for the 500th and 1000th observations using both the linear probability model and the probit model. Interpret the values.

d. Construct a histogram of *FICO*. Using both linear probability and probit models, calculate the probability of default for *FICO* = 500, 600, and 700 for a loan of \$250,000 (*AMOUNT* = 25). For the other variables, the loan to value ratio (*LTV*) is 80%, initial interest rate is 8%, indicator variables take the value 0 except for *TERM30* = 1. Discuss similarities and differences among the predicted probabilities from the two models.

e. Using both linear probability and probit models, compute the marginal effect of *FICO* on the probability of delinquency for *FICO* = 500, 600, and 700, given that the other explanatory variables take the values in (d). Discuss the interpretation of the marginal effect.

f. Construct a histogram of *LTV*. Using both linear probability and probit models, calculate the probability of delinquency for *LVR* = 20 and *LVR* = 80, with *FICO* = 600 and other variables set as they are in (d). Compare and contrast the results.

g. Compare the percentage of correct predictions from the linear probability model and the probit model using a predicted probability of 0.5 as the threshold.

h. As a loan officer, you wish to provide loans to customers who repay on schedule and are not delinquent. Suppose you have available to you the first 500 observations in the data on which to base your loan decision on the second 500 applications (501-1,000). Is using the probit model with a threshold of 0.5 for the predicted probability the best decision rule for deciding on loan applications? If not, what is a better rule? [*Note*: for *vegas5* use the first 5000 observations for the estimation sample and the second 5000 observations for prediction.]

16.20 This exercise deals with the loan data in the data file *lasvegas* described in Exercise 6.18. The "Chow" test was introduced in Section 7.2.3 for testing the equality of coefficients in two regressions on subsets of observations. Here we ask a similar question concerning the parameters of the logit model for delinquency for the two subpopulations of borrowers who either have mortgage insurance ($INSUR = 1$) or not ($INSUR = 0$).

a. Using all observations, estimate the logit model for *DELINQUENT* using all explanatory variables except *INSUR*. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_R$.

b. Reestimate the model in (a) using the sample observations for which $INSUR = 0$. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_0$.

c. Reestimate the model in (b) using the sample observations for which $INSUR = 1$. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_1$.

d. Compare the estimates from the models in (a–c). What major differences in coefficient signs, magnitudes, and significance do you observe?

e. Reestimate the model in (a) including each explanatory variable, as well as *INSUR*, and its interactions with all the other variables. Compare the value of the log-likelihood function from the fully interacted model, call it $\ln L_U$, to $\ln L_0 + \ln L_1$. If you have done things correctly, then $\ln L_U$ should equal $\ln L_0 + \ln L_1$. Can you explain why this must be so?

f. Carry out a likelihood ratio version of the Chow test by computing $LR = 2(\ln L_U - \ln L_R)$. What is the appropriate critical value for a test at the 5% level of significance? What conclusion do you draw about the subgroups of individuals who do and do not have mortgage insurance? Do the two groups behave in the same way?

16.21 Data on 1500 purchases of canned lite tuna are in the data file *tunafish*. There are four brands of tuna (Starkist – water, Starkist – oil, Chicken of the Sea – water, Chicken of the Sea – oil). The A.C. Nielsen data were made available through the University of Chicago's Graduate School of Business. The data file *tunafish_small* is a smaller dataset with 250 purchases. The data are in "stacked" format with four data lines per purchase, one for each tuna brand. The consumer choice is indicated by the indicator variable *CHOICE*. Relevant variables are *NETPRICE* = price minus coupon value, if used; $DISPLAY = 1$ if product is on display, $FEATURE = 1$ if item is featured, and *INCOME* = household income.

a. What is the primary variable-type distinction between *NETPRICE* and *INCOME*?

b. What is the sample percentage of purchases for each brand? What do you observe about consumer preferences for these product choices?

c. Using the conditional logit model, write the probability of choosing each brand using as explanatory variables *NETPRICE*, *DISPLAY*, and *FEATURE*, plus an alternative specific constant using Starkist packed in water as the base category.

d. Estimate the model specified in part (c).

e. For the model in (d) find the marginal effect of *NETPRICE* on the probability of choice of each brand, using for all brands $DISPLAY = FEATURE = 1$. Do these marginal effects have the signs you anticipate? Are the marginal effects statistically significant?

f. Add the variable *INCOME* to the model specified in (c). Perform a likelihood ratio test of its significance.

g. For the model in (f) find the marginal effect of *NETPRICE* on the probability of choice of each brand, using for all brands $DISPLAY = FEATURE = 1$ and $INCOME = 30$.

16.22 How do age, education, and other personal characteristics predict our assessment of our health status? Use the data file *rwm88* to answer the following.

a. Tabulate the values of the variable *HSAT3*, which is a self-rating of health satisfaction, with 1 being the lowest and 3 being highest. What percentages fall into each of the health status categories?

b. Estimate an ordered probit model predicting *HSAT3* using *AGE*, AGE^2, *EDUC2* = years of education, *FEMALE* = 1 if female, *MARRIED* = 1 if married, *HHKIDS* =1 if there are children under age 16 in the household, and *WORKING* = 1 if employed, 0 otherwise. Which variables have coefficients that are statistically significant at the 5% level?

c. Estimate the probability that an employed, unmarried male, age 40 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2.

d. Estimate the probability that an employed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2.

e. Estimate the probability that an employed, unmarried male, age 40 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3.

f. Estimate the probability that an employed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3.

g. Estimate the probability that an unemployed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2. Compare this probability to the result in part (d).

h. Estimate the probability that an unemployed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3. Compare this probability to the result in part (f).

16.23 How well do age, education, and other personal characteristics predict our assessment of our health status? Use the data file *rwm88* to answer the following.

a. Tabulate the variable *HSAT3*, which is a self-rating of health satisfaction, with 1 being the lowest and 3 being highest. What percent of the sample assess their health status as *HSAT3* = 1, 2, or 3?

b. Estimate an ordered probit model predicting *HSAT3* using *AGE*, AGE^2, *EDUC2* = years of education, and *WORKING* = 1 if employed, 0 otherwise. Which variables have coefficients that are statistically significant at the 5% level?

c. Estimate the marginal impact of age on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 40, with 16 years of education, and who is working.

d. Estimate the marginal impact of age on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 70, with 16 years of education, and who is working.

e. Estimate the marginal impact of *WORKING* on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 40, with 16 years of education.

16.24 Consider household expenditures per person on apparel. Use the data file *cex5* for this exercise.

a. What percentage of the households spent nothing on apparel in the previous quarter?

b. Estimate a linear regression with *APPAR* as dependent variable and use as explanatory variables *INCOME*, *SMSA* (*S*tandard *M*etropolitan *S*tatistical *A*rea = 1 if household lives in an urban area, and = 0 otherwise), *ADVANCED*, *COLLEGE*, and *OLDER* (= 1 if someone in the household is 65 years of age or older). Discuss the signs and significance of the estimated coefficients. Interpret the coefficient of *INCOME*. Interpret the coefficient of *ADVANCED*.

c. Repeat the estimation in (b) using only observations for which *APPAR* > 0. What are your answers to the questions in (b) now?

d. Create the variable *SHOP* = 1 if *APPAR* > 0, and *SHOP* = 0 otherwise. Estimate a probit model with dependent variable *SHOP* as a function of the variables in (b). What factors significantly affect the decision to buy clothing?

e. Estimate a Tobit model with dependent variable *APPAR*. Compare the coefficient estimates signs and significance to those in (b) and (c). Calculate the marginal effect of income on the expected amount spent on *APPAREL* for a household living in an urban area, with income $65,000, containing someone with an advanced degree and no one 65 or older in the household. Repeat the calculation for a household with $125,000 income.

16.25 Consider using data file *mroz* to estimate a model explaining a married woman's hours of work, *HOURS*, as a function of her education, *EDUC*, her experience, *EXPER*, and her husband's hours of work, *HHOURS*.

a. Use all observations to estimate the regression model

$$HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HHOURS + e$$

Is OLS a consistent estimator in this case?

b. Use only the observations for which $HOURS > 0$ to estimate the regression model in (a). Is OLS a consistent estimator in this case?

c. Estimate a probit model for the woman's decision to be in the labor force, $LFP = 1$, $LFP = \Phi(\gamma_1 + \gamma_2 EXPER + \gamma_3 KIDSL6 + \gamma_4 KIDS618 + \gamma_5 MTR + \gamma_6 LARGECITY)$. Which if any of the variables help explain the woman's labor force participation decision?

d. Using the estimates from the probit model, obtain

$$\tilde{w} = \tilde{\gamma}_1 + \tilde{\gamma}_2 EXPER + \tilde{\gamma}_3 KIDSL6 + \tilde{\gamma}_4 KIDS618 + \tilde{\gamma}_5 MTR + \tilde{\gamma}_6 LARGECITY$$

Create the inverse Mills ratio $\tilde{\lambda} = \phi(\tilde{w})/\Phi(\tilde{w})$. What are the sample mean and variance of $\tilde{\lambda}$?

e. Estimate the model $HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HHOURS + \beta_\lambda \tilde{\lambda} + e$ using the observations for which $HOURS > 0$. Compare these estimates to those in parts (a) and (b). Are the standard errors from this estimation correct?

f. Estimate the model in (e) using heteroskedasticity robust standard errors. Use the option HC3 if it is available. These standard errors are not absolutely correct but an improvement over the ones in (e).

g. Estimate the model in (e) using bootstrap standard errors, with $B = 400$ bootstrap replications. Compare these standard errors to those in (e) and (f).

h. Estimate the model in (e) using proper econometric software for this Heckit model. Compare the results to those in (e)–(g). Be sure to identify whether your software is using a two-step estimator, like part (e), or full information maximum likelihood.

16.26 In Example 7.11, we used the linear probability model to check whether students were assigned randomly to small classes in Project STAR. In this exercise, we use multinomial logit and the data file *star* to explore the issue.

a. Create the variable $CLASS = 1$ for a regular sized class, $CLASS = 2$ for a small class, and $CLASS = 3$ for a regular sized class with a teacher aide. What percentage of the students in the sample were assigned to each type of class?

b. Estimate a multinomial logit model explaining *CLASS* with explanatory variables *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, *SCHURBAN*, and *SCHRURAL*. Use $CLASS = 1$, the regular class, as the base group. If students are assigned randomly what values should the model coefficients take? Are any of the estimated coefficients significantly different from zero at the 5% level?

c. Find the ratio of the probability of being in a small class for a white boy who receives lunch if his school is in a rural area, relative to the probability of him being in a regular sized class.

d. Find the ratio of the probability of being in a regular sized class with a teacher aide for a white boy who receives lunch if his school is in a rural area, relative to the probability of him being in a regular sized class.

e. Carry out a likelihood ratio test that the coefficients of *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, and *SCHURBAN* are zero, against the alternative that they are not, at the 5% level. What is the 5% critical value for this test?

f. Carry out a likelihood ratio test that the coefficients of *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, *SCHURBAN,* and *SCHRURAL* are zero, against the alternative that they are not, at the 5% level. What is the 5% critical value for this test?

g. Based on the outcomes of parts (a)–(f), what do you conclude about random assignment of students in Project STAR?

16.27 In Example 16.15, we considered a count data model for the number of doctor visits by an individual as a function of a few explanatory variables. In this exercise, we expand the analysis using a larger data set in the data file, *rwm88*, and more explanatory variables. Adjust the data in the following ways: (i) omit individuals for whom $HHNINC2 = 0$; (ii) create the variable $LINC = ln(HHNINC2)$; (iii) create $AGE2 = AGE^2$; (iv) create the variable $POST = 1$ (a postsecondary degree indicator variable) if $FACHHS = 1$ or if $UNIV = 1$, and $POST = 0$ otherwise.

a. Using the first 3000 observations estimate a Poisson model explaining *DOCVIS* as a function of *FEMALE*, *AGE*, *AGE2*, *SELF*, *LINC*, *POST*, and *PUBLIC*. Discuss the signs and the significance of the coefficients on *FEMALE*, *SELF*, *POST*, and *PUBLIC*. Calculate the percentage increase in the expected number of doctor visits for each factor represented by these indicator variables.

b. Compute the estimated percentage change in the expected number of doctor visits associated with another year of age for a person who is 30 years old; who is 50 years old; and who is 70 years old.

c. Interpret the estimated coefficient of *LINC*.

d. Calculate the expected number of doctor visits for each person, *EDOCVIS*, and round this value to the nearest integer to obtain *NVISITS*, the predicted number of visits for each person. Create a variable that indicates a successful prediction. Let *SUCCESS* = 1 if *NVISITS* = *DOCVIS* and *SUCCESS* = 0 otherwise. What is the percentage of successful predictions for observations 1–3000? What is the percentage of successful predictions for the remaining 979 observations?

e. Create *SUCCESS1* which indicates a successful prediction of more than one doctor visit. That is, create a variable *DOCVIS1* = 1 if an individual has more than one doctor visit, and *PREDICT1* = 1 if the model has predicted more than one doctor visit. Let *SUCCESS1* = 1 if *DOCVIS1* = *PREDICT1* and *SUCCESS1* = 0 otherwise. What is the percentage of successful predictions of more than one doctor visit for observations 1–3000? What is the percentage of successful predictions of more than one doctor visit for the remaining 979 observations?

16.28 We have used Ray Fair's voting data, (data file *fair5*, throughout the book to predict presidential election outcomes with the linear regression model. Here we apply probit to predict the outcome of the 2016 U.S. Presidential election. Create the variable *DEMWIN* = 1 if *VOTE* $\geq$ 50.0 and *DEMWIN* = 0 otherwise. As of October 28, 2016, the values for the key economic variables were *GROWTH* = 0.97, *INFLAT* = 1.42, and *GOODNEWS* = 2.

a. Estimate a probit model for *DEMWIN* as a function of *GROWTH*, *INFLAT*, *GOODNEWS* using data for years prior to 2016. Comment on the signs and significance of the estimated coefficients.

b. Using the probit model in part (a), and the given values of *GROWTH*, *INFLAT*, and *GOODNEWS*, predict the election outcome in 2016. What is the estimated probability that a democrat will win?

c. Add *DPER*, *DUR*, *WAR*, and *INCUMB* to the model used in (a). Reestimate the probit model. What happens to the signs and significance of the estimated coefficients?

d. Using the model in (c), obtain the estimated probability, *PHAT*, of a democrat winning for the sample period 1916–2012. Are any of the predicted values very close to 1.0 or 0.0? For how many observations is *PHAT* > 0.99999? For how many observations is *PHAT* < 0.00001?

e. Examine the values of *DEMWIN* when the following four-variable pattern exists in the data: *DPER* = −1, *DUR* = 0, *WAR* = 0, *INCUMB* = −1. How many such observations are there? [*Note*: Some software will indicate probit failure when the dependent variable does not vary for a value of an independent variable, or in this case a particular combination of values. You may think of this as something like "perfect collinearity." When this happens maximum likelihood estimation including the particular pattern of observations fails.]

16.29 In this exercise, we illustrate some features of instrumental variables estimation, and two-stage least squares, when the potential endogenous variable is binary. Use the data file *rwm88* for this problem, and do not worry too much about the economic reasoning behind the model.

a. Estimate by OLS the regression of *DOCVIS* on *AGE*, *FEMALE*, *WORKING*, *HHNINC2*, and *ADDON*. Use heteroskedasticity robust standard errors. Does it appear that having add-on insurance is a significant factor affecting the number of doctor visits?

b. *ADDON* might be endogenous. Estimate a first stage equation using OLS with *ADDON* as dependent variable and *AGE*, *FEMALE*, *WORKING*, *HHNINC2*, *WHITEC*, and *SELF* as explanatory variables. Since the dependent variable is binary use heteroskedasticity robust standard errors. Are *WHITEC* and *SELF* jointly significant? Why does this matter if our objective is two-stage least squares estimation?

c. Obtain the fitted value from part (b), $\widehat{ADDON}$, and reestimate the model in (a) using $\widehat{ADDON}$ in place of *ADDON*. Use heteroskedasticity robust standard errors. Does it appear that having add-on insurance is a significant factor affecting the number of doctor visits?

d. Use your software command designed for two-stage least squares and estimate the model in (a) using external instruments *WHITEC* and *SELF*. Use heteroskedasticity robust standard errors. How do these estimates compare to those in part (c)? Has two-stage least squares performed as expected?

e. Since *ADDON* is binary, estimate the first stage equation in (b) using probit. Compute the estimated probability that *ADDON* = 1, *PHAT*. Reestimate the model in (a) using *PHAT* in place of *ADDON*. Use heteroskedasticity robust standard errors. Are the results the same as in part (d)? Why not?

f. Use your software command designed for two-stage least squares and estimate the model in (a) using external instrument *PHAT*. Use heteroskedasticity robust standard errors. How do these estimates compare to those in part (e)? Has two-stage least squares performed as expected?

16.30 In this exercise, we use multinomial logit to describe factors leading an individual to fall into one of three categories. Use data file *rwm88* for this exercise.

a. Create a variable called *INSURED* = 1, if a person does not have public insurance or add-on insurance (*PUBLIC* = 0 and *ADDON* = 0). Let *INSURED* = 2 if (*PUBLIC* = 1 and *ADDON* = 0). Let *INSURED* = 3 if (*PUBLIC* = 1 and *ADDON* = 1). Tabulate the number of individuals falling into each category. How many individuals are accounted for?

b. Estimate a multinomial logit model with outcome variable *INSURED* and explanatory variables *AGE*, *FEMALE*, *WORKING*, and *HHNINC2*. Use *INSURED* = 1 as the base category. What information is provided by the signs and significance of the estimated coefficients?

c. Obtain the predicted probabilities of falling into each category for each person in the sample, calling them *P1*, *P2*, and *P3*. Find the sample averages of *P1*, *P2*, and *P3* and compare these to the percentages of the sample for whom *INSURED* = 1, 2, and 3, respectfully.

d. Obtain the predicted probabilities of falling into each category for a person who is 50 years old, female, working and with a household income, *HHNINC2* = 2400.

e. Repeat the calculations in (d) for *HHNINC2* = 4200.

f. Calculate the 25th and 75th percentiles of *HHNINC2*. Comment on the changes in probabilities computed in parts (d) and (e).

Appendix 16A Probit Marginal Effects: Details

16A.1 Standard Error of Marginal Effect at a Given Point

Consider the probit model $p = \Phi(\beta_1 + \beta_2 x)$. The marginal effect of a continuous x, evaluated at a specific point $x = x_0$, is

$$\left.\frac{dp}{dx}\right|_{x=x_0} = \phi(\beta_1 + \beta_2 x_0)\beta_2 = g(\beta_1, \beta_2)$$

The estimator of the marginal effect is $g(\tilde{\beta}_1, \tilde{\beta}_2)$, where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the maximum likelihood estimators of the unknown parameters. The variance of this estimator was developed in Appendix 5B.2, in (5B.4), and is given by

$$\begin{aligned}\operatorname{var}\left[g(\tilde{\beta}_1, \tilde{\beta}_2)\right] \cong & \left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}\right]^2 \operatorname{var}(\tilde{\beta}_1) + \left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\right]^2 \operatorname{var}(\tilde{\beta}_2) \\ & + 2\left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}\right]\left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\right] \operatorname{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \qquad (16A.1)\end{aligned}$$

The variances and covariances of the estimators come from maximum likelihood estimation. The essence of these calculations is given in Appendix C.8.2. To implement the delta method, we require the derivative

$$\begin{aligned}\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1} &= \frac{\partial\left[\phi(\beta_1 + \beta_2 x_0)\beta_2\right]}{\partial \beta_1} \\ &= \left\{\frac{\partial \phi(\beta_1 + \beta_2 x_0)}{\partial \beta_1} \times \beta_2\right\} + \phi(\beta_1 + \beta_2 x_0) \times \frac{\partial \beta_2}{\partial \beta_1} \\ &= -\phi(\beta_1 + \beta_2 x_0) \times (\beta_1 + \beta_2 x_0) \times \beta_2\end{aligned}$$

The second line above uses the product rule, Derivative Rule 6. To obtain the final result, we used $\partial\beta_2/\partial\beta_1 = 0$ and

$$\begin{aligned}\frac{\partial\phi(\beta_1+\beta_2x_0)}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\beta_1+\beta_2x_0)^2}\right]\\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\beta_1+\beta_2x_0)^2}\left(2\times-\frac{1}{2}\times(\beta_1+\beta_2x_0)\right)\\ &= -\phi(\beta_1+\beta_2x_0)\times(\beta_1+\beta_2x_0)\end{aligned}$$

The second step uses Derivative Rule 7 for exponential functions. Using similar steps, we obtain the other key derivative,

$$\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_2} = \phi(\beta_1+\beta_2x_0)\left[1-(\beta_1+\beta_2x_0)\times\beta_2x_0\right]$$

From the maximum likelihood estimation results using the transportation data example, we obtain the estimator variances and covariances[32]

$$\begin{bmatrix}\widehat{\text{var}}(\tilde\beta_1) & \widehat{\text{cov}}(\tilde\beta_1,\tilde\beta_2)\\ \widehat{\text{cov}}(\tilde\beta_1,\tilde\beta_2) & \widehat{\text{var}}(\tilde\beta_2)\end{bmatrix} = \begin{bmatrix}0.1593956 & 0.0003261\\ 0.0003261 & 0.0105817\end{bmatrix}$$

The derivatives must be evaluated at the maximum likelihood estimates. For the transportation data used in Examples 16.4 and 16.5 for $DTIME = 2$ ($x_0 = 2$), the calculated values of the derivatives are

$$\widehat{\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_1}} = -0.055531 \quad \text{and} \quad \widehat{\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_2}} = 0.2345835$$

Using (16A.1), and carrying out the required multiplication, we obtain the estimated variance and standard error of the marginal effect

$$\widehat{\text{var}}\left[g(\tilde\beta_1,\tilde\beta_2)\right] = 0.0010653 \quad \text{and} \quad \text{se}\left[g(\tilde\beta_1,\tilde\beta_2)\right] = 0.0326394$$

16A.2 Standard Error of Average Marginal Effect

Consider the probit model $p = \Phi(\beta_1+\beta_2x)$. For the transportation data example, the explanatory variable $x = DTIME$. The average marginal effect of this continuous variable is

$$AME = \frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1+\beta_2DTIME_i)\beta_2 = g_2(\beta_1,\beta_2)$$

The estimator of the average marginal effect is $g_2(\tilde\beta_1,\tilde\beta_2)$. To apply the delta method to find $\text{var}\left[g_2(\tilde\beta_1,\tilde\beta_2)\right]$, we require the derivatives

$$\begin{aligned}\frac{\partial g_2(\beta_1,\beta_2)}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left[\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1+\beta_2DTIME_i)\beta_2\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta_1}\left[\phi(\beta_1+\beta_2DTIME_i)\beta_2\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_1}\end{aligned}$$

[32] Using minus the inverse matrix of second derivatives.

The term $\dfrac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}$ we evaluated in the previous section. Similarly, the derivative

$$\begin{aligned}\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2} &= \frac{\partial}{\partial \beta_2}\left[\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1 + \beta_2 DTIME_i)\beta_2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \beta_2}\left[\phi(\beta_1 + \beta_2 DTIME_i)\beta_2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\end{aligned}$$

For the transportation data, we compute

$$\widehat{\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}} = -0.00185 \quad \text{and} \quad \widehat{\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}} = -0.032366$$

Using (16A.1) with g replaced by g_2, and carrying out the required multiplication, we obtain the estimated variance and standard error of the average marginal effect

$$\widehat{\text{var}}\left[g_2(\tilde{\beta}_1, \tilde{\beta}_2)\right] = 0.0000117 \quad \text{and} \quad \text{se}\left[g_2(\tilde{\beta}_1, \tilde{\beta}_2)\right] = 0.003416$$

Appendix 16B Random Utility Models

Economics is a general theory of choice behavior. Individuals make choices that maximize their wellbeing, or welfare, or, as economists term it, "utility." Observers cannot measure utility directly, and we cannot compare the utility, or satisfaction, that Jane enjoys while eating ice cream to Bill's satisfaction. But when a person is confronted with two or more choices, we assume that they make the choice that maximizes their welfare, however that might be defined. If a person must choose between taking a bus to work or driving to work, then, after considering the various costs and benefits, the person's choice reveals their utility maximizing outcome. We can imagine that the utility they receive depends on the attributes of the alternatives. As modelers we can select some such attributes as explanatory variables, but we must recognize that we will never truly understand choices completely; there is a random unexplained component, or random error, in any model.

Choice models, both binary and multinomial, as well as other limited dependent variable models, are often developed using a random utility model framework. Utility, or satisfaction, is unobservable and consequently it is called a **latent variable**, one that must be present but which is unseen. We will illustrate this approach to modeling by developing the probit model of binary choice in the random utility framework.

16B.1 Binary Choice Model

Assume that an individual must choose between two alternatives. Let U_{i1} be the utility derived from alternative one and let U_{i0} be the utility derived from alternative two. Let z_{i1} be the attributes of alternative one as perceived by the ith individual, and let z_{i0} be the attributes of alternative two as perceived by the ith individual. Let w_i represent the attributes of the ith individual. There may be several attributes of the alternatives that are relevant, and several individual characteristics that matter as well, but for simplicity, we will assume that there is but one attribute of each alternative and one individual characteristic. Then, a linear random utility model for each alternative is

$$\begin{aligned}U_{i1} &= \alpha_1 + z_{i1}\delta + w_i\gamma_1 + e_{i1} \\ U_{i0} &= \alpha_0 + z_{i0}\delta + w_i\gamma_0 + e_{i0}\end{aligned} \tag{16B.1}$$

In each model, there is a random error component, e_{i1} and e_{i0}. Assuming strict exogeneity, $E(e_{i1}|z_{i1}, z_{i0}, w_i) = 0$ and the same for e_{i0}, we can write

$$U_{i1} = E(U_{i1}|\cdot) + e_{i1} \quad \text{and} \quad U_{i0} = E(U_{i0}|\cdot) + e_{i0}$$

so that the utility from each part consists of a systematic part and a random part, as we are used to. Each of the expected utility terms is conditional, but we suppress the notation for convenience. Also, note that the individual characteristics w_i have coefficients that are unique to each alternative but that the attributes of alternatives, z_{i1} and z_{i0}, have a common parameter, δ. The logic of this specification will become clear soon.

As in equation (16.1), let the outcome variable be

$$y_i = \begin{cases} 1 & \text{if alternative one is chosen} \\ 0 & \text{if alternative two is chosen} \end{cases} \tag{16B.2}$$

Based on our model of random utility, alternative one will be chosen, and $y_i = 1$, if $U_{i1} \geq U_{i0}$, or if $U_{i1} - U_{i0} \geq 0$, where

$$\begin{aligned} U_{i1} - U_{i0} &= E(U_{i1}|\cdot) + e_{i1} - \left[E(U_{i0}|\cdot) + e_{i0}\right] \\ &= (\alpha_1 - \alpha_0) + (z_{i1} - z_{i0})\delta + w_i(\gamma_1 - \gamma_0) + (e_{i1} - e_{i0}) \end{aligned} \tag{16B.3}$$

The left-hand side variable $U_{i1} - U_{i0}$ is unobservable, but we know the difference in utilities determines an individual's choice. Let $y_i^* = U_{i1} - U_{i0}$ denote the latent variable which is the difference in utilities. Observe what would happen if the characteristics of the individual had the same coefficient in the random utility models (16B.1). Then the individual characteristics would fall out of (16B.3) and would have no effect on the choice. Equation (16B.3) becomes a regression specification by writing it as

$$\begin{aligned} y_i^* &= (\alpha_1 - \alpha_0) + (z_{i1} - z_{i0})\delta + w_i(\gamma_1 - \gamma_0) + (e_{i1} - e_{i0}) \\ &= \beta_1 + \beta_2(z_{i1} - z_{i0}) + \beta_3 w_i + e_i \\ &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \end{aligned} \tag{16B.4}$$

We observe $y_i = 1$ if $y_i^* = U_{i1} - U_{i0} \geq 0$. The probability of an individual choosing alternative one is

$$\begin{aligned} p(\mathbf{x}_i) &= P(y_i = 1|\cdot) = P(y_i^* \geq 0|\cdot) = P\left[(U_{i1} \geq U_{i0})|\cdot\right] \\ &= P\left[E(U_{i1}|\cdot) + e_{i1} \geq E(U_{i0}|\cdot) + e_{i0}\right] \\ &= P\left[e_{i0} - e_{i1} \leq E(U_{i1}|\cdot) - E(U_{i0}|\cdot)\right] \\ &= P\left[e_{i0} - e_{i1} \leq \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right] \\ &= F(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}) \end{aligned} \tag{16B.5}$$

In the last line of (16B.5), $F(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3})$ is the cumulative distribution function of the random variable $e_{i0} - e_{i1}$. In Section 16.2, we used the *cdf* as a convenient device for keeping the probabilities between zero and one, but here it arises quite naturally from the random utility framework.

16B.2 Probit or Logit?

In binary choice problems, economists tend to use probit over logit. The reason follows from assumptions about the random utility models. Suppose that $e_{i1} \sim N(0, \sigma_1^2)$, $e_{i0} \sim N(0, \sigma_0^2)$, and $\text{cov}(e_{i1}, e_{i0}) = \sigma_{10}$. Then $(e_{i0} - e_{i1}) \sim N(0, \sigma^2 = \sigma_0^2 + \sigma_1^2 - 2\sigma_{10})$. Then

$$
\begin{aligned}
p(\mathbf{x}_i) &= P(y_i = 1|\cdot) \\
&= P\left[e_{i0} - e_{i1} \le \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right] \\
&= P\left[\frac{e_{i0} - e_{i1}}{\sigma} \le \frac{\beta_1}{\sigma} + \frac{\beta_2}{\sigma} x_{i2} + \frac{\beta_3}{\sigma} x_{i3}\right] \\
&= \Phi\left(\beta_1^* + \beta_2^* x_{i2} + \beta_3^* x_{i3}\right)
\end{aligned}
$$

The parameters in the probit model are actually $\beta_k^* = \beta_k/\sigma$. The parameter scaling is usually ignored in notation with the explanation that we choose $\sigma = 1$ as a normalization.[33] Then the probit model is $p(\mathbf{x}_i) = \Phi\left(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right)$.

On the other hand, to obtain a logit model, the random errors e_{i1} and e_{i0} must be statistically independent and identically distributed with an *extreme value distribution*.[34] In this case, $\left(e_{i0} - e_{i1}\right) = v_1$ has a logistic distribution. The details are a fun exercise (see Example B.7 for part of it) and outlined in Dhrymes (1986, page 1574).[35]

The bottom line is that there is no reason to assume that the random utility errors are statistically independent, nor to have the asymmetrical extreme value distribution. It is a mathematically convenient assumption because the end result, the logistic distribution, has a *cdf* of convenient form. Assuming that the random utility errors are normally distributed, and correlated, is not at all a stretch of the imagination.

Appendix 16C Using Latent Variables

Using latent variables, we can develop a variety of models that involve observed and partially observed variables. We will illustrate a few using simple models. Others can be found in Amemiya (1984, "Tobit Models: A Survey," *Journal of Econometrics*, 24, pages 3–61).

16C.1 Tobit (Tobit Type I)

Amemiya called the standard Tobit model "Type I Tobit." Let $y_i^* = \beta_1 + \beta_2 x_i + e_i$ be a latent variable with $e_i \sim N\left(0, \sigma^2\right)$. The Tobit model then arises by specifying the observed outcome value y_i to be,

$$
y_i = \begin{cases} y_i^* = \beta_1 + \beta_2 x_i + e_i & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \le 0 \end{cases}
$$

Three possible regression functions are then

$$
E\left(y_i^*|x_i\right) = \beta_1 + \beta_2 x_i
$$

$$
E\left(y_i|x_i, y_i > 0\right) = \beta_1 + \beta_2 x_i + \frac{\phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right]}{\Phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right]}
$$

$$
E\left(y_i|x_i\right) = \Phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right] E\left(y_i|x_i, y_i > 0\right)
$$

[33] The issue of this normalization comes into play in the discussion of Heckman's two-step estimator, discussed in Section 16.7.5.

[34] https://en.wikipedia.org/wiki/Gumbel_distribution

[35] http://www.sciencedirect.com/science/handbooks/15734412/3

The marginal effects for a continuous variable x_i are

$$\partial E(y_i^*|x_i)/\partial x_i = \beta_2$$
$$\partial E(y_i|x_i, y_i > 0)/\partial x_i = \left\{1 - \alpha_i\lambda(\alpha_i) - [\lambda(\alpha_i)]^2\right\}\beta_2$$
$$\partial E(y_i|x_i)/\partial x_i = \Phi(\alpha_i)\beta_2$$

where $\alpha_i = (\beta_1 + \beta_2 x_i)/\sigma$ and $\lambda(\alpha_i) = \phi(\alpha_i)/\Phi(\alpha_i)$.

16C.2 Heckit (Tobit Type II)

The famous model of self-selection (Tobit Type II) developed by James Heckman is called "Heckit." In this model, there are two equations. The selection equation, that describes a person's participation decision, and an intensity, or amount, equation, which is the equation of interest. In the latent variable formulation, the equations are

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i \quad \text{selection equation}$$
$$y_i^* = \beta_1 + \beta_2 x_i + e_i \quad \text{amount equation, the equation of interest}$$

The equations are connected through their error terms. Let $u_i \sim N(0, \sigma_u^2)$ and $e_i \sim N(0, \sigma_e^2)$, with the covariance between these two random errors being σ_{ue}. The latent variables z_i^* and y_i^* are not observed. We do observe the binary variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$y_i = \begin{cases} y_i^* = \beta_1 + \beta_2 x_i + e_i & \text{if } z_i^* > 0 \text{ or } z_i = 1 \\ 0 & \text{if } z_i^* \le 0 \text{ or } z_i = 0 \end{cases}$$

Using a theorem about bivariate normal random variables, similar to Appendix B.3.5, it can be shown that

$$E(y_i|x_i, w_i, y_i > 0) = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi\left[(\gamma_1 + \gamma_2 w_i)/\sigma_u\right]}{\Phi\left[(\gamma_1 + \gamma_2 w_i)/\sigma_u\right]} = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi(\gamma_1^* + \gamma_2^* w_i)}{\Phi(\gamma_1^* + \gamma_2^* w_i)}$$

Heckman's two-step estimator first estimates the selection model's scaled parameters $\gamma_1^* = \gamma_1/\sigma_u$ and $\gamma_2^* = \gamma_2/\sigma_u$ by probit using all observations. Then, using *only positive* observations, estimates by OLS the equation of interest

$$y_i = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi(\tilde{\gamma}_1^* + \tilde{\gamma}_2^* w_i)}{\Phi(\tilde{\gamma}_1^* + \tilde{\gamma}_2^* w_i)} + v_i$$

The two-step estimator is consistent and asymptotically normally distributed, but the usual OLS standard errors are incorrect. The corrected ones are complicated but available in econometric software. An alternative is to estimate by maximum likelihood the two equations jointly, which is a more efficient estimation option. The MLE is often the default in econometric software, so check your documentation.

Appendix 16D A Tobit Monte Carlo Experiment

Let the latent variable be

$$y_i^* = \beta_1 + \beta_2 x_i + e_i = -9 + x_i + e_i \tag{16D.1}$$

with the error term assumed to have a normal distribution, $e_i \sim N(0, \sigma^2 = 16)$. The observable outcome y_i takes the value zero if $y_i^* \leq 0$, but $y_i = y_i^*$ if $y_i^* > 0$. In the simulation, we

- Create $N = 200$ random values of x_i that are spread evenly (or uniformly) over the interval [0, 20].
- Obtain $N = 200$ random values e_i from a normal distribution with mean 0 and variance 16.
- Create $N = 200$ values of the latent variable $y_i^* = -9 + x_i + e_i$.
- Obtain $N = 200$ values of the observed y_i using

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases}$$

The 200 observations obtained this way constitute a sample that is **censored** with a lower limit of zero. The latent data are plotted in Figure 16D.1. In this figure, the line labeled $E(y_i^*|x_i)$ has intercept -9 and slope 1. The values of the latent variable y_i^* (triangle and hollow circle, △ and ○) are scattered along this regression function; if we observed these data we could estimate the parameters using the least squares principle, by fitting a line through the center of the data.

However, we do not observe all the latent data. When the values of y_i^* are zero or less then we observe $y_i = 0$ (●). We observe the y_i^* when they are positive. These observable data, along with the fitted least squares regression, are shown in Figure 16D.2.

The least squares principle will fail to estimate $\beta_1 = -9$ and $\beta_2 = 1$ because the observed data do not fall along the underlying regression function $E(y^*|x) = \beta_1 + \beta_2 x = -9 + x$.

To illustrate, the results from the first Monte Carlo sample, data file *tobit5*, are contained in Table 16D.1. In the first column (y^*) are the OLS estimates using the simulated latent data. In the second column $(y > 0)$ are the OLS estimates using only the 118 observations for which the

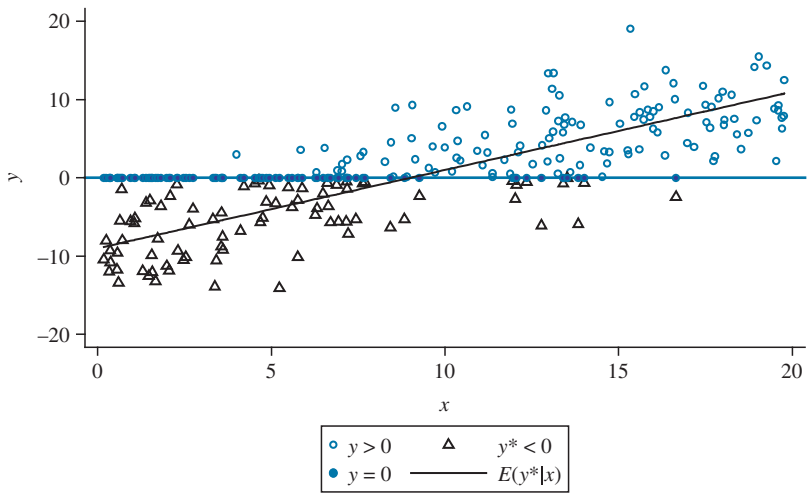


FIGURE 16D.1 **Latent and censored data.**

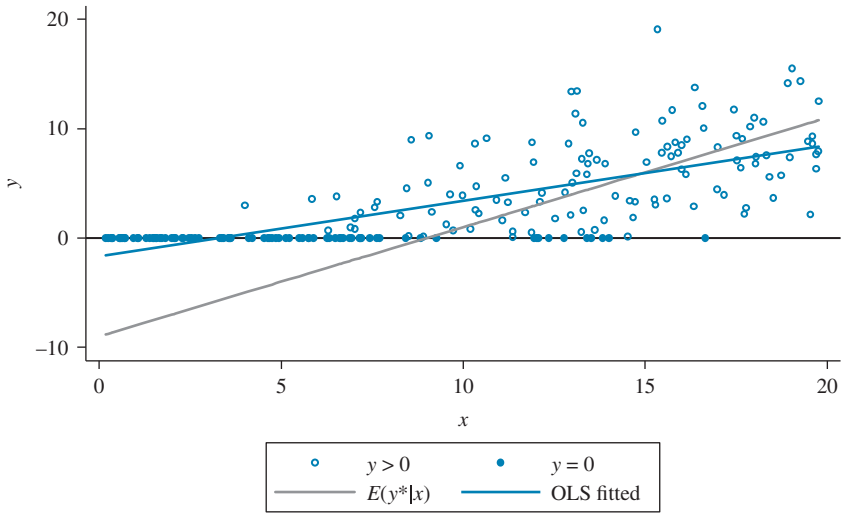


FIGURE 16D.2 **Observed data and OLS fitted line.**

TABLE 16D.1 **Simulated Censored Data (*tobit5*)**

	y^*	$y > 0$	y	Tobit
C	−8.6611	−1.1891	−1.6515	−8.0007
	(0.5842)	(1.1777)	(0.4290)	(0.9802)
x	0.9690	0.5176	0.5075	0.9215
	(0.0499)	(0.0823)	(0.0366)	(0.0722)
$\hat{\sigma}$	4.1050	3.4340	3.0146	3.9884
				(0.2670)
N	200	118	200	200

(Standard errors in parentheses)

observed value of y is positive; in the third column (y), are the OLS estimates on the 200 observed values of y, and in the fourth column are the Tobit estimates. The Tobit estimates are relatively close to the true value, while the estimates based only on the positive y values, or on all the y values, are far from the mark. An added benefit of the ML method is that there is a standard error for the estimated value of σ.

In the Monte Carlo simulation, we repeat this process of creating $N = 200$ observations, and applying least squares estimation, many times. This is analogous to "repeated sampling" in the context of experimental statistics. In this case, we repeat the process $NSAM = 1000$ times, drawing new x-values and error values e, recording each time the values of the estimates we obtain. At the end, we can compute the average values of the estimates which is the Monte Carlo "expected value,"

$$E_{MC}(b_k) = \frac{1}{NSAM}\sum_{m=1}^{NSAM} b_{k(m)}$$

where $b_{k(m)}$ is the estimate of β_k in the mth Monte Carlo sample. We also compute the Monte Carlo average of the usual, or "nominal" standard error, and the standard deviation of the estimates. The standard deviation measures the true sampling variability of the estimates. It is our hope that the usual standard error captures the actual sampling variation so that the average nominal standard error and the standard deviation of the estimates are close. The results are in Table 16D.2.

TABLE 16D.2 **Monte Carlo Simulation Results**

	Intercept = −9			Slope = 1		
	Mean	Standard Error	Standard Deviation	Mean	Standard Error	Standard Deviation
y^*	−9.0021	0.5759	0.5685	1.0000	0.0498	0.0492
$y > 0$	−2.1706	0.9518	1.1241	0.6087	0.0729	0.0779
y	−2.2113	0.2928	0.4185	0.5632	0.0389	0.0362
Tobit	−9.0571	1.0116	0.9994	1.0039	0.0740	0.0733

The results of applying OLS to the latent data (y^*) produce estimates that are on average very close to the true values for both the intercept and the slope. The average of the nominal standard error is close to the standard deviation of the estimates. If we discard the $y = 0$ observations and apply least squares to just the positive y observations, $y > 0$, these averages are −2.1706 and 0.6087, respectively. If we apply the least squares estimation procedure to all the observed censored data (y, including observations $y = 0$), the average value of the estimated intercept is −2.2113, and the average value of the estimated slope is 0.5632. The least squares estimates are biased by a substantial amount, compared to the true values $\beta_1 = -9$ and $\beta_2 = 1$. This bias will not disappear no matter how large the sample size we consider because the least squares estimators are inconsistent when data are censored or truncated. On the other hand, the Tobit estimates on average are very close to the true values.

A word of caution is in order about commercial software packages. There are many algorithms available for obtaining maximum likelihood estimates, and different packages use different ones, which may lead to slight differences (in perhaps the 3rd or 4th decimal) in the parameter estimates and their standard errors. When carrying out important research, it is a good tip to confirm empirical results with a second software package, just to be sure that they give essentially the same numbers.

Appendix A

Mathematical Tools

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Explain the relationship between exponential functions and natural logarithms.
2. Explain and apply scientific notation.
3. Define a linear relationship, as opposed to a nonlinear relationship.
4. Compute the elasticity at a point on a function.
5. Explain the concept of a derivative and its relationship to the slope of a function.
6. Compute the derivatives of simple functions and provide their interpretations.
7. Describe the relationship between a derivative and a partial derivative.
8. Explain the concept of an integral.
9. Maximize or minimize functions of one or two variables.
10. Use integration to find the area under curves.
11. Explain and evaluate second derivatives.

KEYWORDS

absolute value
antilogarithm
derivatives
e
elasticity
exponential function
exponents
inequalities
integers
integral
intercept
irrational number
linear relationship
logarithm
marginal effect
maximizing a function
minimizing a function
natural logarithm
nonlinear relationship
partial derivative
percentage change
product rule
quadratic function
quotient rule
rational numbers
real numbers
relative change
scientific notation
second derivative
slope
Taylor series

We assume that you have studied basic math. Hopefully you understand the calculus concepts of differentiation and integration, though these tools are *not required* prerequisites for success using this book. In this appendix we review some essential concepts that you may wish to consult from time to time.[1]

[1] Summation signs and operations are covered in the Probability Primer that precedes Chapter 2.

A.1 Some Basics

A.1.1 Numbers

Integers are the whole numbers, 0, ±1, ±2, ±3, ... The positive integers are the counting numbers. **Rational numbers** can be written as a/b, where a and b are integers and $b \neq 0$. The **real numbers** can be represented by points on a line. There are an uncountable number of real numbers, and they are not all rational. Numbers such as $\pi \cong 3.1415927$ and $\sqrt{2}$ are said to be **irrational** since they cannot be expressed as ratios, and have only decimal representations. Numbers like $\sqrt{-2}$ are not real numbers. The **absolute value** of a number is denoted by $|a|$. It is the positive part of the number: $|3| = 3$ and $|-3| = 3$.

Inequalities among numbers obey certain rules. The notation $a < b$, a is less than b, means that a is to the left of b on the number line, and that $b - a > 0$. If a is less than or equal to b, it is written as $a \leq b$. Three basic rules are

$$\text{If } a < b, \text{ then } a + c < b + c$$

$$\text{If } a < b, \text{ then } \begin{cases} ac < bc & \text{if } c > 0 \\ ac > bc & \text{if } c < 0 \end{cases}$$

$$\text{If } a < b \text{ and } b < c, \text{ then } a < c$$

A.1.2 Exponents

Exponents are defined as follows:

$x^n = xx \cdots x$ (n terms) if n is a positive integer

$x^0 = 1$ if $x \neq 0$. 0^0 does not have meaning and is "undefined."

Some common rules for working with exponents, assuming x and y are real, m and n are integers, and a and b are rational, are as follows:

$$x^{-n} = \frac{1}{x^n} \text{ if } x \neq 0. \text{ For example, } x^{-1} = \frac{1}{x}$$

$$x^{1/n} = \sqrt[n]{x}. \text{ For example, } x^{1/2} = \sqrt{x} \text{ and } x^{-1/2} = \frac{1}{\sqrt{x}}$$

$$x^{m/n} = \left(x^{1/n}\right)^m. \text{ For example, } 8^{4/3} = \left(8^{1/3}\right)^4 = 2^4 = 16$$

$$x^a x^b = x^{a+b}, \ \frac{x^a}{x^b} = x^{a-b}$$

$$\left(\frac{x}{y}\right)^a = \frac{x^a}{y^a}, \ (xy)^a = x^a y^a$$

A.1.3 Scientific Notation

Scientific notation is useful for very large or very small numbers. A number in scientific notation is written as a number between 1 and 10 multiplied by a power of 10. So, for example: $5.1 \times 10^5 = 510,000$, and $0.00000034 = 3.4 \times 10^{-7}$. Scientific notation makes handling large numbers much easier, because complex operations can be broken into simpler ones. For example,

$$\begin{aligned} 510{,}000 \times 0.00000034 &= \left(5.1 \times 10^5\right) \times \left(3.4 \times 10^{-7}\right) \\ &= (5.1 \times 3.4) \times \left(10^5 \times 10^{-7}\right) \\ &= 17.34 \times 10^{-2} \\ &= 0.1734 \end{aligned}$$

and

$$\frac{510{,}000}{0.00000034} = \frac{5.1 \times 10^5}{3.4 \times 10^{-7}} = \frac{5.1}{3.4} \times \frac{10^5}{10^{-7}} = 1.5 \times 10^{12}$$

Computer programs sometimes write $5.1 \times 10^5 = 5.1\text{E}5$ or $5.1\text{D}5$ and $3.4 \times 10^{-7} = 3.4\text{E}{-7}$ or $3.4\text{D}{-7}$.

A.1.4 Logarithms and the Number *e*

Logarithms are exponents. If $x = 10^b$, then b is the **logarithm** of x using the base 10. The **irrational number** $\boldsymbol{e} \cong 2.718282$ is used in mathematics and statistics as the base for logarithms. If $x = e^b$, then b is the logarithm of x using the base e. Logarithms using the number e as base are called **natural logarithms**. *All logarithms in this book are natural logarithms*. We express the natural logarithm of x as $\ln(x)$,

For any positive number, $x > 0$,

$$e^{\ln(x)} = \exp\left[\ln(x)\right] = x$$

and

$$\ln\left(e^x\right) = x$$

Note that $\ln(1) = 0$, using the laws of exponents. Table A.1 gives the logarithms of some powers of 10. For example, $e^{2.3025851} = 10$ and $e^{4.6051702} = 100$.

Note that logarithms have a compressed scale compared to the original numbers. Since logarithms are exponents, they follow similar rules:

$$\ln(xy) = \ln(x) + \ln(y)$$
$$\ln(x/y) = \ln(x) - \ln(y)$$
$$\ln\left(x^a\right) = a\ln(x)$$

For example, if $x = 1000$ and $y = 10,000$, then

$$\begin{aligned}\ln(1000 \times 10{,}000) &= \ln(1000) + \ln(10{,}000)\\ &= 6.9077553 + 9.2103404\\ &= 16.118096\end{aligned}$$

What is the advantage of this? The value of xy is a multiplication problem, which by using logarithms we can turn into an addition problem. We need a way to go backward, from the logarithm of a number to the number itself. By definition,

$$x = e^{\ln(x)} = \exp\left[\ln(x)\right]$$

TABLE A.1 Some Natural Logarithms

x	$\ln(x)$
1	0
10	2.3025851
100	4.6051702
1,000	6.9077553
10,000	9.2103404
100,000	11.512925
1,000,000	13.815511

When there is an **exponential function** with a complicated exponent, the notation **exp** is often used, so that $e^{(\cdot)} = \exp(\cdot)$. The exponential function is the **antilogarithm**, because we can recover the value of x using it. Then,

$$1000 \times 10000 = \exp(16.118096) = 10{,}000{,}000$$

You will not be doing many calculations like these, but the knowledge of logarithms and exponents is quite critical in economics and econometrics.

A.1.5 Decimals and Percentages

Suppose the value of a variable y changes from the value $y = y_0$ to $y = y_1$. The difference between these values is often denoted by $\Delta y = y_1 - y_0$, where the notation Δy is read "change in y," or "delta-y." The **relative change in** y is defined to be

$$\text{relative change in } y = \frac{y_1 - y_0}{y_0} = \frac{\Delta y}{y_0} \tag{A.1}$$

For example, if $y_0 = 3$ and $y_1 = 3.02$, then the relative change in y is

$$\frac{y_1 - y_0}{y_0} = \frac{3.02 - 3}{3} = 0.0067$$

Often the relative change in y is written as $\Delta y/y$, omitting the subscript.

A relative change is a decimal. The corresponding **percentage change** in y is 100 times the relative change.

$$\text{percentage change in } y = 100\frac{y_1 - y_0}{y_0} = \%\Delta y \tag{A.2}$$

If $y_0 = 3$ and $y_1 = 3.02$, then the percentage change in y is

$$\%\Delta y = 100\frac{y_1 - y_0}{y_0} = 100\frac{3.02 - 3}{3} = 0.67\%$$

A.1.6 Logarithms and Percentages

A feature of logarithms that helps greatly in their economic interpretation is that they can be approximated very simply. Let y_1 be a positive value of y, and let y_0 be a value of y that is "close" to y_1. A useful approximation rule is

$$100\left[\ln(y_1) - \ln(y_0)\right] \cong \%\Delta y = \text{percentage change in } y \tag{A.3}$$

That is, 100 times the difference in the logarithms is the approximate percentage difference between y_0 and y_1, if y_0 and y_1 are close.

Derivation of the Approximation The result in (A.3) follows from the mathematical tool called a **Taylor series** approximation, which is developed in Example A.3 in Section A.3.1. Using this approximation, the value of $\ln(y_1)$ can be written as

$$\ln(y_1) \cong \ln(y_0) + \frac{1}{y_0}(y_1 - y_0) \tag{A.4}$$

For example, let $y_1 = 1 + x$ and let $y_0 = 1$. Then, as long as x is small,

$$\ln(1 + x) \cong x$$

Subtracting $\ln(y_0)$ from both sides of (A.4), we obtain

$$\ln(y_1) - \ln(y_0) = \Delta\ln(y) \cong \frac{1}{y_0}(y_1 - y_0) = \text{relative change in } y$$

The symbol $\Delta\ln(y)$ represents the "difference" between two logarithms. Using (A.2),

$$\begin{aligned} 100\Delta\ln(y) &= 100\left[\ln(y_1) - \ln(y_0)\right] \\ &\cong 100 \times \frac{(y_1 - y_0)}{y_0} = \%\Delta y \\ &= \text{percentage change in } y \end{aligned}$$

A.2 Linear Relationships

In economics, and in econometrics, we study linear and nonlinear relationships between variables. In this section, we review basic characteristics of **linear relationships**. Let y and x be variables. The standard form for a linear relationship is

$$y = mx + b \tag{A.5}$$

In Figure A.1, the **slope** is m and the y-**intercept** is b. The symbol Δ represents "a change in," so Δx is read as a "change in x." The slope of the line is

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

For the straight-line relationship in Figure A.1, the slope m is the ratio of the change in vertical distance (rise) to the change in horizontal distance (run) as a point moves along the line in either direction. The slope of a straight line is constant; the rate at which y changes as x changes is constant over the length of the straight line.

The slope m is very meaningful to economists as it is the **marginal effect** of a change in x on y. To see this, solve the **slope** definition $m = \Delta y/\Delta x$ for Δy, obtaining

$$\Delta y = m\Delta x \tag{A.6}$$

If x changes by one unit, $\Delta x = 1$, then the corresponding change in y is $\Delta y = m$. The marginal effect, m, is always the same for a linear relationship like (A.5), because the slope is constant.

The **intercept** parameter indicates where the linear relationship crosses the vertical axis—that is, it is the value of y when x is zero,

$$y = mx + b = m \times 0 + b = b$$

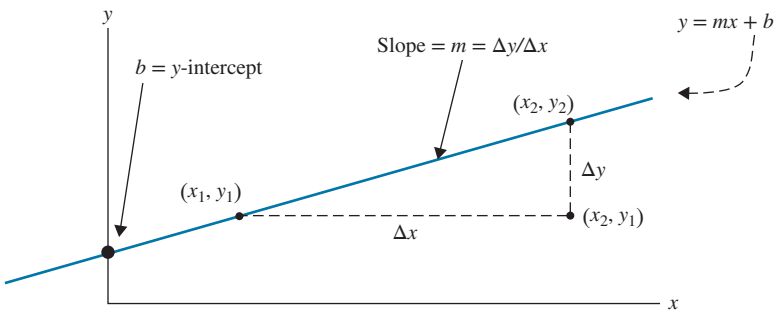


FIGURE A.1 **A linear relationship.**

A.2.1 Slopes and Derivatives

Derivatives have an important role in econometrics. In a relationship between two variables, $y = f(x)$, the **first derivative** measures the slope. The slope of the line $y = f(x) = mx + b$ is denoted as dy/dx. The notation dy/dx is a "stylized" version of $\Delta y/\Delta x$, and for the linear relationship (A.5) the first derivative is

$$dy/dx = m \tag{A.7}$$

In general, the first derivative measures the change in the function value y given an infinitesimal change in x. For the linear function the first derivative is the constant $m = \Delta y/\Delta x$. The "infinitesimal" does not matter in this case, because the rate of change of y with respect to changes in x is a constant.

A.2.2 Elasticity

A favorite tool of the economist is **elasticity**. It is the percentage change in one variable associated with a 1% change in another variable for movements along a specific curve. That is, if we move from one point on a curve to another point on the curve, what are the relative percentage changes? For example, in Figure A.1, what is the percentage change in y relative to the percentage change in x as we move from the point (x_1, y_1) to (x_2, y_2)? For a linear relationship, the elasticity of y with respect to a change in x is

$$\varepsilon_{yx} = \frac{\%\Delta y}{\%\Delta x} = \frac{100(\Delta y/y)}{100(\Delta x/x)} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x} \times \frac{x}{y} = slope \times \frac{x}{y} \tag{A.8}$$

The elasticity is the product of the slope of the relationship and the ratio of an x value to a y value. In a linear relationship, such as Figure A.1, while the slope is constant, $m = \Delta y/\Delta x$, the elasticity changes at every (x, y) point on the line.

Consider, for example, the linear function $y = 1x + 1$. At the point $x = 2$ and $y = 3$, which is on the line, the elasticity is $\varepsilon_{yx} = m(x/y) = 1 \times (2/3) = 0.67$. That is, at the point $(x = 2, y = 3)$ a 1% change in x is associated with a 0.67% change in y. Specifically, at $x = 2$ a 1% (1% = 0.01 in decimal form) change is $\Delta x = 0.01 \times 2 = 0.02$. If x increases to $x = 2.02$, the value of y increases to 3.02. The **relative change** in y is $\Delta y/y = 0.02/3 = 0.0067$. This, however, is not the percentage change in y, but rather the decimal equivalent. To obtain the percentage change in y, which we denote $\%\Delta y$, we multiply the relative change $\Delta y/y$ by 100. The **percentage change** in y is

$$\%\Delta y = 100 \times (\Delta y/y) = 100 \times 0.02/3 = 100 \times 0.0067 = 0.67\%$$

A.3 Nonlinear Relationships

While linear relationships are intuitive and easy to work with, many real-world economic relationships are nonlinear, as illustrated in Figure A.2.

The slope of this curve is not constant. The slope measures the marginal effect of x on y, and for a **nonlinear relationship** like that in Figure A.2, the slope is different at every point on the curve. The changing slope tells us that the relationship is not linear. Since the slope is different at every point, we can only talk about the effect of small changes in x on y. In (A.6) we replace Δ, the symbol for "a change in," with d, which we will take to mean an "infinitesimal change in." In the linear case when we made this replacement, the slope was given by $dy/dx = m$, where m was a constant. See equation (A.7).

However, with nonlinear functions such as that in Figure A.2, the slope (derivative) is not constant, but changes as x changes, and must be determined at each point. Strictly speaking, the slope of a curve is the slope of the **tangent** to the curve at a specific point. To work out the slope at different points on a nonlinear curve, we need some rules for obtaining the derivative dy/dx.

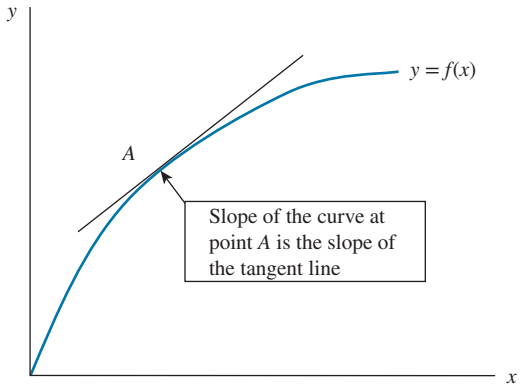


FIGURE A.2 **A nonlinear relationship.**

A.3.1 Rules for Derivatives

Some rules for finding derivatives are the following:

Derivative Rule 1. The derivative of a constant c is zero, that is, if $y = f(x) = c$, then

$$\frac{dy}{dx} = 0$$

Derivative Rule 2. If $y = x^n$, then

$$\frac{dy}{dx} = nx^{n-1}$$

Derivative Rule 3. If $y = cu$ and $u = f(x)$, then

$$\frac{dy}{dx} = c\frac{du}{dx}$$

Constants can be factored out of functions before taking the derivative.

Derivative Rule 4. If $y = cx^n$, using Rules 2 and 3,

$$\frac{dy}{dx} = cnx^{n-1}$$

Derivative Rule 5. If $y = u + v$, where $u = f(x)$ and $v = g(x)$ are functions of x, then

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx}$$

The derivative of the sum (or difference) of two functions is the sum (or difference) of the derivatives. This rule extends to more than two terms in a sum.

Derivative Rule 6. If $y = uv$, where $u = f(x)$ and $v = g(x)$ are functions of x, then

$$\frac{dy}{dx} = \frac{du}{dx}v + u\frac{dv}{dx}$$

This is called the **product rule**. The **quotient rule**, for $y = u/v$, is obtained by inserting v^{-1} for v in the product rule.

Derivative Rule 7. If $y = e^x$, then

$$\frac{dy}{dx} = e^x$$

If $y = \exp(ax + b)$, then

$$\frac{dy}{dx} = \exp(ax + b) \times a$$

In general, the derivative of the exponential function is the exponential function times the derivative of the exponent.

Derivative Rule 8. If $y = \ln(x)$, then

$$\frac{dy}{dx} = \frac{1}{x}, \quad x > 0$$

If $y = \ln(ax + b)$, then

$$\frac{dy}{dx} = \frac{1}{ax + b} \times a$$

Derivative Rule 9. (The Chain Rule of Differentiation). Let $y = f[u(x)]$, so that y depends on u which in turn depends on x. Then

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$$

For example, in Derivative Rule 8, $y = \ln(ax + b)$, or $y = \ln[u(x)]$ where $u = ax + b$. Then

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx} = \frac{1}{u} \times a = \frac{1}{ax + b} \times a$$

EXAMPLE A.1 | Slope of a Linear Function

The derivative of $y = f(x) = 4x + 1$ is

$$\frac{dy}{dx} = \frac{d(4x)}{dx} + \frac{d(1)}{dx} = 4$$

Because this function is the equation of a straight line, $y = mx + b$, its slope is constant and given by the coefficient of x, which in this case is 4.

EXAMPLE A.2 | Slope of a Quadratic Function

Consider the function $y = x^2 - 8x + 16$, shown in Figure A.3. This **quadratic function** is a parabola. Using the rules of derivatives, the slope of a line tangent to the curve is

$$\frac{dy}{dx} = \frac{d(x^2 - 8x + 16)}{dx} = \frac{d(x^2)}{dx} - 8\frac{d(x^1)}{dx} + \frac{d(16)}{dx}$$

$$= 2x^1 - 8x^0 + 0 = 2x - 8$$

This result means that the slope of the tangent line to this curve is $dy/dx = 2x - 8$. The derivative and function values are shown for several values of x in Table A.2.

Note a few things. First, the slope is different at each value of x. The slope is negative for values of $x < 4$, the slope is zero when $x = 4$, and the slope is positive for values of $x > 4$. To interpret these slopes, recall that the derivative of a function at a point is the slope of the tangent at that point. The slope of the tangent is the **rate of change** of the function—how much $y = f(x)$ is changing as x changes. At $x = 0$, the derivative is -8, indicating that y is falling as x increases, and that the rate of change is 8 units in y per unit change in x. At $x = 2$, the rate of change of the function has diminished, and at $x = 4$, the rate of change of the function is $dy/dx = 0$. That is, at $x = 4$, the slope of the tangent to the curve is zero. For values of $x > 4$, the derivative is positive, which indicates that the function $y = f(x)$ is increasing as x increases.

TABLE A.2 **The Function $y = x^2 - 8x + 16$ and Derivative Values**

x	$y = f(x)$	dy/dx
0	16	−8
2	4	−4
4	0	0
6	4	4
8	16	8

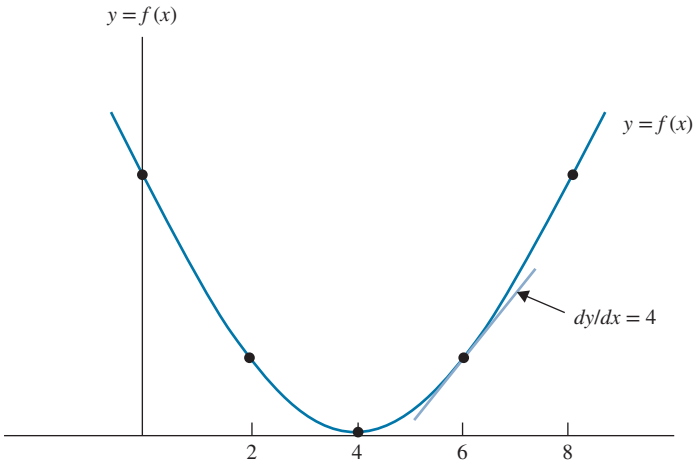


FIGURE A.3 **The function $y = x^2 - 8x + 16$.**

EXAMPLE A.3 | Taylor Series Approximation

The approximation of the logarithm in (A.4) uses a very powerful tool called a Taylor series approximation. For the function $f(y) = \ln(y)$ it is illustrated in Figure A.4. Assume that we know the point A on the function: for $y = y_0$, we know the function value $f(y_0) = \ln(y_0)$. The approximation idea is to draw a line tangent to the curve $f(y) = \ln(y)$ at A, then approximate the point on the curve $f(y_1) = \ln(y_1)$ by the point B on the tangent line. For a smooth curve like $\ln(y)$, this strategy works well, and the approximation error will be small if y_1 is close to y_0. The slope of the tangent line at point A, $(y_0, f(y_0) = \ln(y_0))$, is the derivative of the function $f(y) = \ln(y)$ evaluated at y_0. Using Derivative Rule 8, we have

$$\left.\frac{d\ln(y)}{dy}\right|_{y=y_0} = \left.\frac{1}{y}\right|_{y=y_0} = \frac{1}{y_0}$$

The value of the linear approximation at B is given by geometry. Recall that the slope of the tangent (straight) line is "the

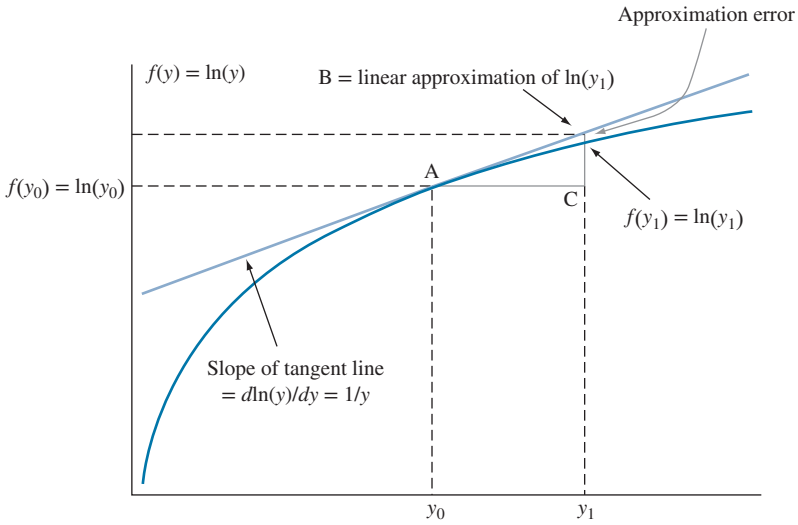


FIGURE A.4 **Taylor series approximation of ln(*y*).**

rise over the run." The "run" is A to C, or $(y_1 - y_0)$, and the corresponding "rise" is C to B. Then

$$\text{tangent slope} = \left.\frac{d\ln(y)}{dy}\right|_{y=y_0} = \frac{1}{y_0} = \frac{\text{rise}}{\text{run}} = \frac{\overline{\text{CB}}}{\overline{\text{AC}}} = \frac{\text{B} - \ln(y_0)}{y_1 - y_0}$$

Solving this equation for B = approximate value of $f(y_1)$, we obtain the expression in (A.4),

$$\text{B} = \ln(y_0) + \left.\frac{d\ln(y)}{dy}\right|_{y=y_0}(y_1 - y_0) = \ln(y_0) + \frac{1}{y_0}(y_1 - y_0)$$

The Taylor series approximation is used in many contexts.

Derivative Rule 10. (Taylor series approximation). If $f(x)$ is a smooth function, then

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x - a) = f(a) + f'(a)(x - a)$$

where $f'(a)$ is a common notation for the first derivative of the function $f(x)$ evaluated at $x = a$. The approximation is good for x close to a. See Exercise A.16 for a **second-order Taylor series approximation**.

A.3.2 Elasticity of a Nonlinear Relationship

Given the slope of a curve, the elasticity of y with respect to changes in x is given by a slightly modified (A.8),

$$\varepsilon_{yx} = \frac{dy/y}{dy/x} = \frac{dy}{dx} \times \frac{x}{y} = slope \times \frac{x}{y}$$

For example, the quadratic function $y = ax^2 + bx + c$ is a parabola. The slope (derivative) is $dy/dy = 2ax + b$. The elasticity is

$$\varepsilon_{yx} = slope \times \frac{x}{y} = (2ax + b)\frac{x}{y}$$

As a numerical example, consider the curve defined by $y = f(x) = x^2 - 8x + 16$. The graph of this quadratic function is shown in Figure A.3. The slope of the curve is $dy/dx = 2x - 8$. When $x = 6$, the slope of the tangent line is $dy/dx = 4$. When $x = 6$, the corresponding value of $y = 4$. So the elasticity at that point is

$$\varepsilon_{xy} = (dy/dx) \times (x/y) = (2x - 8)(x/y) = 4(6/4) = 6$$

A 1% increase in x is associated with a 6% change in y.

A.3.3 Second Derivatives

Since the derivative dy/dx of $f(x)$ is a function of x itself, we can define the derivative of the first derivative of $f(x)$, or **second derivative** of $f(x)$, as

$$\frac{d^2y}{dx^2} = \frac{d(dy/dx)}{dx}$$

The second derivative of a function is interpreted as the rate of change of the first derivative and indicates whether the function is increasing or decreasing at an increasing, constant or decreasing rate.

EXAMPLE A.4 | Second Derivative of a Linear Function

Find the second derivative of $y = 4x + 1$. Using the rules of differentiation

$$\frac{dy}{dx} = \frac{d(4x+1)}{dx} = 4$$

and

$$\frac{d^2y}{dx^2} = \frac{d(dy/dx)}{dx} = \frac{d(4)}{dx} = 0.$$

The function $y = f(x) = 4x + 1$ is a straight line and has a constant first derivative, or slope, 4. The rate of change of the first derivative is zero, and the function increases at a constant rate.

EXAMPLE A.5 | Second Derivative of a Quadratic Function

Find the second derivative of the function $y = x^2 - 8x + 16$ shown in Figure A.3.

$$\frac{dy}{dx} = \frac{d(x^2 - 8x + 16)}{dx} = 2x - 8$$

$$\frac{d^2y}{dx^2} = \frac{d(2x-8)}{dx} = 2$$

The second derivative of $y = f(x)$ is positive and the constant 2, which indicates that the first derivative is increasing for $-\infty < x < \infty$. For $x < 4$ the function is decreasing at a decreasing rate since the negative slope becomes less steep; for $x > 4$ the function increases at an increasing rate. At $x = 4$ the function is at its minimum and the slope is zero.

A.3.4 Maxima and Minima

Using first and second derivatives, we can define relative, or local, maxima and minima of functions, as shown in Figure A.5.

The function $y = f(x)$ has a relative or local maximum at $x = a$ if $f(a)$ is greater than any other value of $f(x)$ in an interval around $x = a$; the function $y = f(x)$ has a relative or local minimum at $x = a$ if $f(a)$ is less than any other value of $f(x)$ in an interval around $x = a$. The conditions for a local maximum or minimum of a function $y = f(x)$ at $x = a$ are as follows:

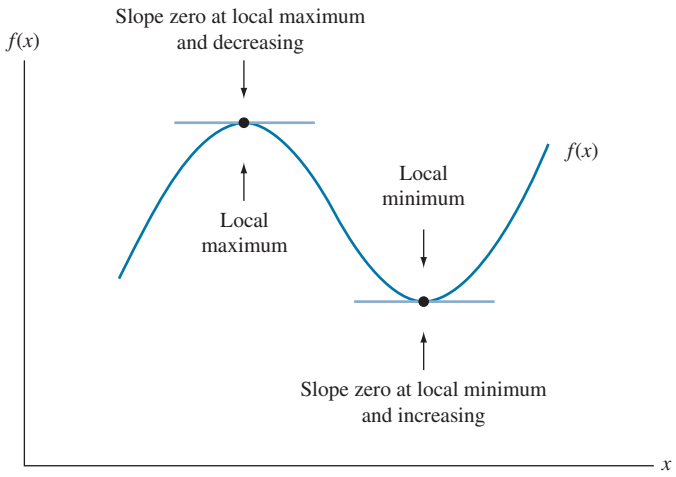


FIGURE A.5 **Local maxima and minima.**

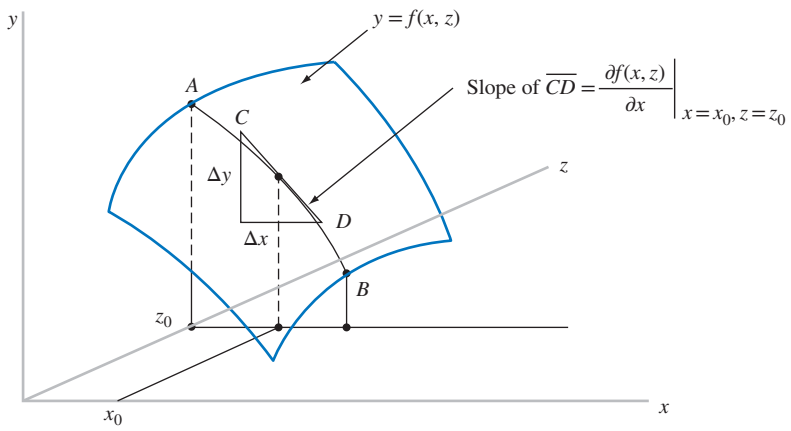


FIGURE A.6 **Three-dimensional diagram of a partial derivative.**

If $y = f(x)$ and dy/dx are nice (continuous) functions at $x = a$, and if $dy/dx = 0$ at $x = a$ then

1. If $d^2y/dx^2 < 0$ at $x = a$ then $f(a)$ is a local maximum.
2. If $d^2y/dx^2 > 0$ at $x = a$ then $f(a)$ is a local minimum.

EXAMPLE A.6 | Finding the Minimum of a Quadratic Function

In Examples A.3 and A.5, we considered the function $y = x^2 - 8x + 16$. To locate possible local minima or maxima, obtain the first derivative, set it to zero, and solve for values of x where $dy/dx = 0$. For this function, $dy/dx = 2x - 8 = 0$ implies that at $x = 4$ we may have a local maximum or a local minimum. Since $d^2y/dx^2 = 2 > 0$, the function is increasing at an increasing rate at $x = 4$ (and everywhere else), and thus $f(4) = 0$ is a local minimum of $y = x^2 - 8x + 16$.

Two notes regarding Example A.6: first, $y = f(x)$ achieves its global or absolute minimum at $x = 4$ as well as its local minimum. Second, if $dy/dx = 0$ at a point $x = a$ where $d^2y/dx^2 = 0$ then the "test" for a local maxima or minima using first and second derivatives does not apply.

A.3.5 Partial Derivatives

When a functional relationship includes several variables, such as $y = f(x, z)$, the slope depends on the values of x and z, and there are slopes in two directions rather than one. In Figure A.6, we illustrate the **partial derivative** of the function with respect to x, holding z constant at the value $z = z_0$.

At the point (x_0, z_0), the value of the function is $y_0 = f(x_0, z_0)$. The slope of the tangent line $\overline{CD}$ is the partial derivative.

$$\text{Slope of } \overline{CD} = \left.\frac{\partial f(x, z)}{\partial x}\right|_{x=x_0, z=z_0}$$

The vertical bar indicates that the partial derivative function is evaluated at the point (x_0, z_0).

To find the partial derivative, we use the already established rules. Consider the function

$$y = f(x, z) = ax^2 + bx + cz + d$$

To find the partial derivative of y with respect to x, treat z as a constant. Then

$$\frac{\partial y}{\partial x} = \frac{d(ax^2)}{dx} + \frac{d(bx)}{dx} + \frac{d(cz)}{dx} + \frac{d(d)}{dx} = 2ax + b$$

Using Derivative Rule 1, the third and fourth terms in the derivative are zero, because cz and d are treated as constants.

A.3.6 Maxima and Minima of Bivariate Functions

Let $y = f(x, z)$ be a continuous function of two variables, or a **bivariate function**, with continuous first derivatives. In order for the point $(x = a, z = b)$ to be a local maximum or minimum three conditions must be met.

1. The two partial derivatives be zero when evaluated at that point:

$$\left.\frac{\partial y}{\partial x}\right|_{x=a,\, z=b} = 0, \qquad \left.\frac{\partial y}{\partial z}\right|_{x=a,\, z=b} = 0$$

   These slope conditions are depicted in Figure A.7.

2. For a local maximum, shown in Figure A.7(a), the second partial derivatives must both be negative at the point $(x = a, z = b)$

$$\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b} < 0, \qquad \left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b} < 0$$

   These two conditions ensure that the function is concave and moving downward in the directions of the x and z axes.

   For a local minimum, shown in Figure A.7(b), the second partial derivatives must both be positive at the point $(x = a, z = b)$ so that the function is convex and the function is moving upward in both the x and z directions

$$\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b} > 0, \qquad \left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b} > 0$$

3. For a local maximum or minimum, the product of the second-order direct partials evaluated at $(x = a, z = b)$ must be larger than the square of the second-order cross-partial derivative at $(x = a, z = b)$, that is,

$$\left(\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b}\right)\left(\left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b}\right) > \left(\left.\frac{\partial^2 y}{\partial x \partial z}\right|_{x=a,\, z=b}\right)^2$$

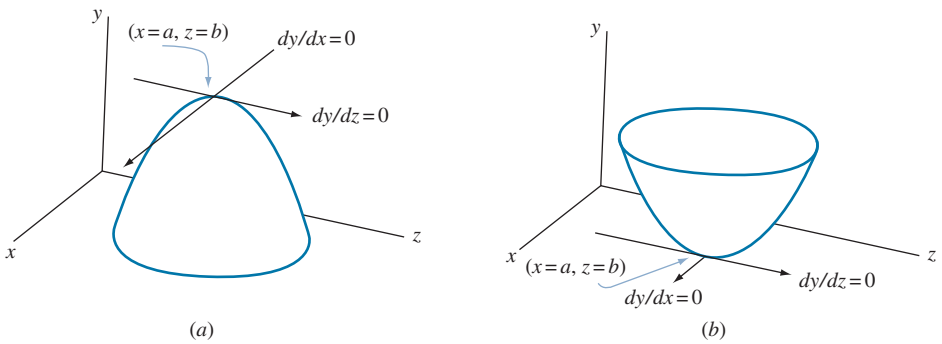


FIGURE A.7 **(*a*) Local maximum and (*b*) local minimum.**

For a local maximum, this condition ensures that the function is moving downward in all directions from $(x = a, z = b)$, not just along the x and z axes. For a local minimum, this condition ensures that the function is moving upward in all directions from $(x = a, z = b)$, not just along the x and z axes.

EXAMPLE A.7 | Maximizing a Profit Function

A firm produces two goods, x and y. The firm's profit function is $\pi = 64x - 2x^2 + 4xy - 4y^2 + 32y - 14$. Find the profit maximizing level of output of x and y. The first partial derivatives are

$$\partial\pi/\partial x = 64 - 4x + 4y, \qquad \partial\pi/\partial y = 4x - 8y + 32$$

The first condition for a maximum or minimum is to set these first derivatives to zero and solve for possible profit maximizing values (x^*, y^*)

$$\left.\begin{aligned} 64 - 4x + 4y &= 0 \\ 4x - 8y + 32 &= 0 \end{aligned}\right\} \Rightarrow x^* = 40, \quad y^* = 24$$

These two values may maximize profit, minimize profit, or neither. We must check the second and third conditions above. The second direct and cross-partial derivatives are

$$\frac{\partial^2\pi}{\partial x^2} = \frac{\partial(64 - 4x + 4y)}{\partial x} = -4$$

$$\frac{\partial^2\pi}{\partial y^2} = \frac{\partial(4x - 8y + 32)}{\partial y} = -8$$

$$\frac{\partial^2\pi}{\partial x\partial y} = \frac{\partial(64 - 4x + 4y)}{\partial y} = 4$$

Both of the second direct partial derivatives are negative, satisfying the second condition for a local maximum. The third condition is that

$$\left(\frac{\partial^2\pi}{\partial x^2}\right)\left(\frac{\partial^2\pi}{\partial y^2}\right) > \left(\frac{\partial^2\pi}{\partial x\partial y}\right)^2$$

This condition is satisfied too, since $(-4)(-8) = 32 > (4)^2 = 16$. Thus, profit is maximized at $x^* = 40$, $y^* = 24$, and the maximum profit is $\pi^* = 1650$.

EXAMPLE A.8 | Minimizing a Sum of Squared Differences

The least squares problem is to find values α and β that minimize the objective function $S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ where (y_i, x_i), $i = 1, \ldots, n$ are data values. Given three pairs of data values $(y_1, x_1) = (1, 1)$, $(y_2, x_2) = (5, 2)$, and $(y_3, x_3) = (2, 3)$, find the minimizing values of α and β.

To find the minimizing values we first expand

$$\begin{aligned} S(\alpha, \beta) &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \\ &= \sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i) \\ &= \sum_{i=1}^{n} y_i^2 + n\alpha^2 + \beta^2\sum_{i=1}^{n} x_i^2 - 2\alpha\sum_{i=1}^{n} y_i - 2\beta\sum_{i=1}^{n} x_i y_i \\ &\quad + 2\alpha\beta\sum_{i=1}^{n} x_i \end{aligned}$$

For the $n = 3$ given data pairs

$$\sum_{i=1}^{3} y_i^2 = 30, \quad \sum_{i=1}^{3} x_i^2 = 14, \quad \sum_{i=1}^{3} y_i = 8,$$

$$\sum_{i=1}^{3} x_i y_i = 17, \quad \sum_{i=1}^{3} x_i = 6$$

The objective function is then

$$\begin{aligned} S(\alpha, \beta) &= 30 + 3\alpha^2 + \beta^2(14) - 2\alpha(8) - 2\beta(17) + 2\alpha\beta(6) \\ &= 30 + 3\alpha^2 + 14\beta^2 - 16\alpha - 34\beta + 12\alpha\beta \end{aligned}$$

The first direct partial derivatives are

$$\frac{\partial S(\alpha, \beta)}{\partial\alpha} = 6\alpha - 16 + 12\beta, \qquad \frac{\partial S(\alpha, \beta)}{\partial\beta} = 28\beta - 34 + 12\alpha$$

Setting these two equations to zero and solving yields $\alpha^* = 5/3$ and $\beta^* = 1/2$. The second-order partial derivatives are

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha^2} = \frac{\partial(6\alpha - 16 + 12\beta)}{\partial\alpha} = 6$$

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\beta^2} = \frac{\partial(28\beta - 34 + 12\alpha)}{\partial\beta} = 28$$

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha\partial\beta} = \frac{\partial(6\alpha - 16 + 12\beta)}{\partial\beta} = 12$$

Both second direct partial derivatives are positive, and the third condition is satisfied because

$$\begin{aligned} \left(\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha^2}\right)\left(\frac{\partial^2 S(\alpha, \beta)}{\partial\beta^2}\right) &= 6(28) \\ &= 168 > \left(\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha\partial\beta}\right)^2 = 144 \end{aligned}$$

Thus, the values $\alpha^* = 5/3$, $\beta^* = 1/2$ minimize the least squares objective function, which takes the value $S(\alpha^*, \beta^*) \cong 8.167$.

A.4 Integrals

An **integral** is an "antiderivative." If $f(x)$ is a function, we can ask the question, "Of what function $F(x)$ is this the derivative?" The answer is given by the **indefinite integral**

$$\int f(x)\,dx = F(x) + C$$

The function $f(x) + C$, where C is a constant called the **constant of integration**, is an antiderivative of $f(x)$ because

$$\frac{d[F(x) + C]}{dx} = \frac{d[F(x)]}{dx} + \frac{d[C]}{dx} = f(x)$$

Finding $F(x)$ is an application of reversing the rules for derivatives. For example, using the rules of derivatives,

$$\frac{d\left(x^n + C\right)}{dx} = nx^{n-1}$$

Thus, $\int nx^{n-1}dx = x^n + C = F(x) + C$, so in this case $F(x) = x^n$. Many indefinite integrals have been worked out and are tabled in your favorite calculus book and at many websites.

Some handy facts about integrals are as follows:

Integral Rule 1.

$$\int \left[f(x) + g(x)\right]dx = \int f(x)dx + \int g(x)dx$$

An integral of a sum is the sum of the integrals.

Integral Rule 2.

$$\int cf(x)dx = c\int f(x)dx$$

Constants can be factored out of integrals.

These rules can be combined so that

Integral Rule 3.

$$\int \left[c_1 f(x) + c_2 g(x)\right]dx = c_1 \int f(x)\,dx + c_2 \int g(x)dx$$

Integral Rule 4 (power rule).

$$\int x^n dx = \frac{1}{n+1}x^{n+1} + C, \quad \text{where } n \neq -1$$

Integral Rule 5 (power rule $n = -1$).

$$\int x^{-1}dx = \ln(x) + C \text{ for } x > 0$$

Integral Rule 6 (constant function).

$$\int k\,dx = kx + C$$

Integral Rule 7 (exponential function).

$$\int e^{kx}dx = \frac{1}{k}e^{kx} + C$$

A.4.1 Computing the Area Under a Curve

An important use of integrals in econometrics and statistics is to calculate areas under curves. For example, in Figure A.8, what is the shaded area under the curve $f(x)$?

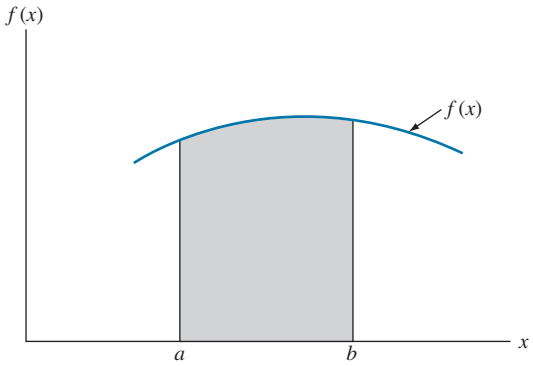


FIGURE A.8 **Area under a curve.**

The area between a curve $f(x)$ and the x-axis, between the limits a and b, is given by the **definite integral**

$$\int_a^b f(x)dx$$

The value of this integral is provided by the **fundamental theorem of calculus**, which says that

$$\int_a^b f(x)dx = F(b) - F(a)$$

EXAMPLE A.9 | Area Under a Curve

Consider the function

$$f(x) = \begin{cases} 2x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(A.12)}$$

This is the equation of a straight line through the origin, as shown in Figure A.9.

What is the shaded area in Figure A.9, the area under the line between a and b? The answer can be found using the geometry of triangles. The area of a triangle is half the base times the height, $\frac{1}{2} \times base \times height$. Triangles can be identified by their corners. Let $\Delta 0bc$ represent the area of the triangle formed by the points 0 (the origin), b, and c. Similarly $\Delta 0ad$ represents the area of the smaller triangle formed by the points 0, a, and d. The shaded area that represents the area under $f(x) = 2x$ between a and b is the difference between the areas of these two triangles.

$$\begin{aligned} \text{Area} &= \Delta 0bc - \Delta 0ad \\ &= \left(\frac{1}{2}b\right)(2b) - \frac{1}{2}a(2a) \\ &= b^2 - a^2 \end{aligned} \qquad \text{(A.13)}$$

Equation (A.13) gives us an easy formula for calculating the area under $f(x) = 2x$ falling between a and b.

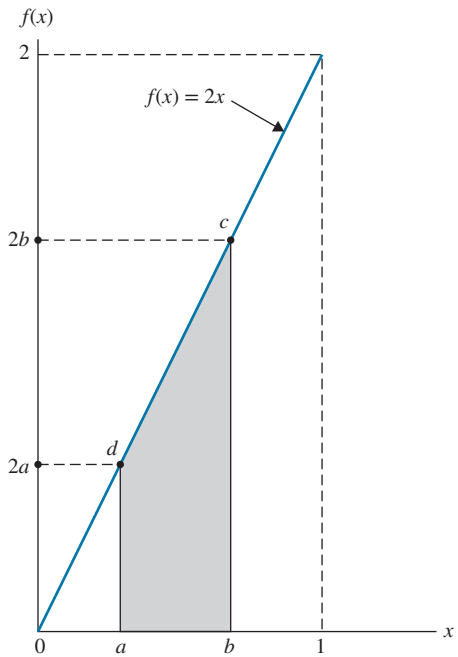


FIGURE A.9 **Area under the curve $f(x) = 2x, 0 \le x \le 1$.**

Using integration, the area under the curve $f(x) = 2x$ and above the x-axis between the limits $x = a$ and $x = b$ is obtained by finding the **definite integral** of $f(x) = 2x$. To use the fundamental theorem of calculus, we need the indefinite integral. Using the power rule, Integral Rule 4, we obtain

$$\int 2xdx = 2\int xdx = 2\left[\frac{1}{2}x^2 + C\right] = x^2 + 2C$$
$$= x^2 + C_1 = F(x) + C_1$$

where $F(x) = x^2$ and the constant of integration is C_1. The area we seek is given by

$$\int_a^b 2xdx = F(b) - F(a) = b^2 - a^2 \qquad \text{(A.14)}$$

This is the same answer we obtained in (A.13) using geometry.

Many times the algebra is abbreviated, because the constant of integration does not affect the definite integral. You will see for definite integrals

$$\int_a^b 2xdx = x^2\Big|_a^b = b^2 - a^2$$

The vertical bar notation means: evaluate the expression first at b and subtract from it the value of the expression at a.

A.5 Exercises

A.1 Each of the following formulas, (1), (2), and (3), represents a supply or demand relation.

(1) $Q = -3 + 2P$ where $P = 10$
(2) $Q = 100 - 20P$ where $P = 4$
(3) $Q = 50P^{-2}$ where $P = 2$

a. Calculate the slope of each function at the given point.
b. Interpret the slope found in (a). Do the slopes change for different values of P and Q? Is it a supply curve (positive relationship) or a demand curve (inverse relationship)?
c. Calculate the elasticity of each function at the given point.
d. Interpret the elasticity found in (c). Do the elasticities change for different values of P and Q?

A.2 The infant mortality rate (*MORTALITY*) for a country is related to the annual per capita income (*INCOME*, U.S. $1000) in that country. Three relationships that may describe this relationship are

(1) $\ln(MORTALITY) = 7.5 - 0.5\ln(INCOME)$
(2) $MORTALITY = 1400 - 100INCOME + 1.67INCOME^2$
(3) $MORTALITY = 1500 - 50INCOME$

a. Sketch each of these relationships between *MORTALITY* and *INCOME* between $INCOME = 0$ and $INCOME = 30$.
b. For each of these relationships, calculate the elasticity of infant mortality with respect to income if (i) $INCOME = 1$, (ii) $INCOME = 3$, and (iii) $INCOME = 25$.

A.3 Suppose the rate of inflation *INF*, the annual percentage increase in the general price level, is related to the annual unemployment rate *UNEMP* by the equation $INF = -3 + 7 \times (1/UNEMP)$.

a. Sketch the curve for values of *UNEMP* between 1 and 10.
b. Where is the impact of a change in the unemployment rate the largest?
c. If the unemployment rate is 5%, what is the marginal effect of an increase in the unemployment rate on the inflation rate?

A.4 Simplify the following expressions:

a. $x^{2/3}x^{2/7}$
b. $x^{2/3} \div x^{2/7}$
c. $\left(x^6y^4\right)^{-1/2}$

A.5 Below are the 2015 *GDP* ($US) figures provided by the World Bank for a few countries.

- **a.** Express each in scientific notation.
  - **i.** Maldives *GDP* $3,142,812,004
  - **ii.** Nicaragua *GDP* $12,692,562,187
  - **iii.** Ecuador *GDP* $100,871,770,000
  - **iv.** New Zealand *GDP* $173,754,075,210
  - **v.** India *GDP* $2,073,542,978,208
  - **vi.** United States *GDP* $17,946,996,000,000
- **b.** Using scientific notation divide the U.S. *GDP* by the *GDP* in (i) Maldives (ii) Ecuador.
- **c.** The population of New Zealand in 2015 was 4.595 million. Use calculations with scientific notation to compute the per capita income in New Zealand. Express the result in scientific notation.
- **d.** The 2015 population of St. Lucia was 184,999 and its *GDP* was $1,436,390,325. Use calculations with scientific notation to compute the per capita income in St. Lucia. Express the result in scientific notation.
- **e.** Using scientific notation, express the sum of the U.S. and New Zealand *GDP* values. [*Hint*: Write each number as $a10^x$ where x is a convenient number for both and a is a numerical value, then simplify.]

A.6 Technology affects agricultural production by increasing yield over time. Let $WHEAT_t$ = average wheat production (tonnes per hectare) for the period 1950–2000 ($t = 1, \ldots, 51$) in Western Australia's Mullewa Shire.

- **a.** Suppose production is defined by $WHEAT_t = 0.58 + 0.14\ln(t)$. Plot this curve. Find the slope and elasticity at the point $t = 49$ (1998).
- **b.** Suppose production is defined by $WHEAT_t = 0.78 + 0.0003\,t^2$. Plot this curve. Find the slope and elasticity at the point $t = 49$ (1998).

A.7 Consider the function $WAGE = f(AGE) = 10 + 200AGE - 2AGE^2$.

- **a.** Sketch the curve for values of *AGE* between $AGE = 20$ and $AGE = 70$.
- **b.** Find the derivative $dWAGE/dAGE$ and evaluate it at $AGE = 30$, $AGE = 50$, and $AGE = 60$. On the curve in part (a), sketch the tangent to the curve at $AGE = 30$.
- **c.** Find the *AGE* at which *WAGE* is maximized.
- **d.** Compute $WAGE_1 = f(29.99)$ and $WAGE_2 = f(30.01)$. Locate these values (approximately) on your sketch from part (a).
- **e.** Evaluate $m = [f(30.01) - f(29.99)]/0.02$. Compare this value to the value of the derivative computed in (b). Explain, geometrically, why the values should be close. The value m is a "numerical derivative," which is useful for approximating derivatives.

A.8 Sketch each of the demand curves below. (i) Indicate the area under the curve between prices $P = 1$ and $P = 2$ on the sketch. (ii) Using integration, calculate the area under the curve between prices $P = 1$ and $P = 2$.

- **a.** $Q = 15 - 5P$
- **b.** $Q = 10P^{-1/2}$
- **c.** $Q = 10/P$

A.9 Consider the function $f(y) = 1/100$ over the interval $0 < y < 100$ and $f(y) = 0$ otherwise.

- **a.** Calculate the area under the curve $f(y)$ for the interval $30 < y < 50$ using a geometric argument.
- **b.** Calculate the area under the curve $f(y)$ for the interval $30 < y < 50$ as an integral.
- **c.** What is a general expression for the area under $f(y)$ over the interval $[a, b]$, where $0 < a < b < 100$?
- **d.** Calculate the integral from $y = 0$ to $y = 100$ of the function $yf(y) = y/100$.

A.10 Consider the function $f(y) = 2e^{-2y}$ for $0 < y < \infty$.

- **a.** Draw a sketch of the function.
- **b.** Compute the integral of $f(y)$ from $y = 1$ to $y = 2$ and illustrate the value on the part (a) sketch.

A.11 Let $y_0 = 1$. For each of the values $y_1 = 1.01, 1.05, 1.10, 1.15, 1.20$, and 1.25 compute

- **a.** The actual percentage change in y using equation (A.2).
- **b.** The approximate percentage change in y using equation (A.3).
- **c.** Comment on how well the approximation in equation (A.3) works as the value of y_1 increases.

A.12 A firm uses labor (L) and capital (K) to produce output (Q). Suppose the production function is $Q = 6L^{1/2}K^{1/3}$. The firm sells its product at price $P = 4$ and pays its labor a wage $W = 12$ with the price of capital being $R = 5$.

- **a.** Find the combination of labor and capital that maximizes profits $\pi = P \times Q - (W \times L) - (R \times K)$ where Q is given by the production function. Check all conditions for a relative maximum.
- **b.** Find the marginal product of labor, $\partial Q/\partial L$, and the marginal product of capital, $\partial Q/\partial K$, at the profit maximizing amounts of labor and capital.

A.13 Use Derivative Rule 10 (Taylor Series approximation) to approximate each of the functions below at $x = 1.5$ and $x = 2$. Let $a = 1$. Calculate the percentage approximation error in each case.

- **a.** $f(x) = 3x^2 - 5x + 1$
- **b.** $f(x) = \ln(2x)$
- **c.** $f(x) = e^{2x}$

A.14 Suppose that a person's earnings (*INCOME*) are determined by their education (*EDUC*) and experience (*EXPER*) according to the relation

$$INCOME = -2EDUC^2 + 78EDUC - 2EXPER^2 + 66EXPER - 2EDUC \times EXPER$$

Find the values of education and experience that maximize the person's income.

A.15 A variable y changes value from $y_0 = 4$ to $y_1 = 4.6$.

- **a.** Compute the relative change in y.
- **b.** Compute the percentage change in y.
- **c.** If the value of y is 4, what is the value of y if it increases by 18%?

A.16 Derivative Rule 10 is a "first-order" Taylor series approximation. A **"second-order" Taylor series approximation** is

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x-a) + \frac{1}{2}\left.\frac{d^2f(x)}{dx^2}\right|_{x=a}(x-a)^2$$

$$= f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

where $f''(a)$ represents the second derivative of the function evaluated at the point $x = a$.

- **a.** Use both the first- and second-order Taylor series approximations to approximate the function $f(x) = e^{2x}$ at $x = 1.5$ and $x = 2$. Let $a = 1$. Calculate the percentage approximation error in each case.
- **b.** Draw a sketch of the function $f(x) = e^{2x}$ for $0 < x < 3$. On the sketch show the tangent line to the function at $a = 1$. On the same graph extrapolate the tangent line to show the location of the first-order approximation when $x = 2$. Show the value of the second-order approximation when $x = 2$.
- **c.** Calculate the percentage approximation error for the first- and second-order Taylor series approximations in part (b). Which is better in this case?

A.17 In 2015, the *GDP* (in nominal U.S. dollars) of Belarus was $GDP_B = \$54,608,962,634.99$ and that of Poland was $GDP_P = \$474,783,393,022.95$.

- **a.** Write GDP_B in scientific notation.
- **b.** Use scientific notation to divide GDP_P by GDP_B. Show your work.
- **c.** Write the natural log of GDP_P.
- **d.** Find $\exp\left[\ln\left(GDP_A\right) - \ln\left(GDP_B\right)\right]$. Write the solution in scientific notation. Show your work.

A.18 Carry out the following:

- **a.** Suppose your wage rate increases from \$17/hr to \$18/hr. What is the percentage increase in your wage?
- **b.** Calculate 100[ln(18) – ln(17)].
- **c.** Suppose your wage rate increases from \$17/hr to \$28/hr. What is the percentage increase in your wage?
- **d.** Calculate 100[ln(28) – ln(17)].
- **e.** Calculate ln(1.02).
- **f.** Calculate ln(1.57).

A.19 Suppose your wage rate is determined by

$$WAGE = -19.68 + 2.52EDUC + 0.55EXPER - 0.007EXPER^2$$

where *EDUC* is years of schooling and *EXPER* is years of work experience. Using calculus, what value of *EXPER* maximizes *WAGE* for a person with 16 years of education? Show your work.

A.20 Suppose wages are determined by the following equation. *EDUC* = years of education, *EXPER* = years of work experience, and *FEMALE* = 1 if person is female, 0 otherwise.

$$WAGE = -23.06 + 2.85EDUC + 0.80EXPER - 0.008EXPER^2 - 9.21FEMALE + 0.34(FEMALE \times EDUC) - 0.015(EDUC \times EXPER)$$

Find $\partial WAGE/\partial EDUC$ for a female with 16 years of schooling and 10 years of experience. Show your work.

Appendix B

Probability Concepts

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Explain the difference between a random variable and its values, and give an example.
2. Explain the difference between discrete and continuous random variables, and give examples of each.
3. State the characteristics of probability density functions (*pdf*) for discrete and continuous random variables, and give examples illustrating these characteristics.
4. Compute probabilities of events, given the probability density function for a discrete or continuous random variable.
5. Show, geometrically and algebraically, using integration, how to compute probabilities given a *pdf* for a continuous random variable.
6. Use the definitions of expected values for discrete and continuous random variables to compute expectations, given a *pdf* $f(x)$ and a function $g(x)$.
7. Define the variance of a random variable, and explain in what sense the values of a random variable are more spread out if the variance is larger.
8. Use a joint *pdf* for two continuous random variables to compute probabilities of joint events, and to find the (marginal) *pdf* of each individual random variable.
9. Find the conditional *pdf* for one random variable given the value of another and their joint *pdf*, and use it to compute conditional probabilities, the conditional mean, and the conditional variance.
10. Define the covariance and correlation between two random variables, and compute these values given a joint probability function.
11. Explain and apply the law of iterated expectations. Explain the variance and covariance decompositions.
12. Find the distribution of a random variable $Y = g(X)$, when $g(X)$ is a strictly increasing or decreasing function, given the probability density function $f(x)$ for the random variable X.
13. Obtain a random number from a probability density function $f(x)$ when its cumulative distribution function $F(x)$ is invertible.
14. Explain in what sense random numbers generated by a computer are random, and in what sense they are not.

KEYWORDS

binary variable
binomial random variable
cdf
change of variable technique
chi-square distribution
conditional *pdf*
conditional probability
continuous random variables
correlation
covariance
covariance decomposition
cumulative distribution function
degrees of freedom
discrete random variable
expected value
experiment
F-distribution
inversion method
iterated expectation
Jacobian
joint probability density function
marginal distributions
mean
median

modulus
normal distribution
pdf
Poisson distribution
probability
probability density function
pseudo-random numbers
random number
random number seed
random variable
standard deviation
standard normal distribution
statistically independent
strictly monotonic
t-distribution
uniform distribution
variance
variance decomposition

We assume that you have had a basic probability and statistics course and that you have read the Probability Primer that precedes Chapter 2. If you have not read the Probability Primer, then do so now.

In this appendix we summarize rules of expected values and variances for **discrete random variables** for easy reference. We then develop similar rules for **continuous random variables** that will require the use of integral concepts introduced in Appendix A.4. We review the properties of some important discrete and continuous random variables, including the *t*-, chi-square, and *F*-distributions. Finally, we introduce concepts related to computer-generated random numbers.

B.1 Discrete Random Variables

In this section we provide a summary of operations with discrete random variables. See the Probability Primer for examples and general background discussion.

A **random variable** is a variable whose value is unknown until it is observed; in other words, it is a variable that is not perfectly predictable. A **discrete random variable** can take only a limited, or countable, number of values. An example of a discrete random variable is the number of late credit card bill payments last year by a *randomly* selected individual. A special case occurs when a random variable can only be one of two possible values. A payment is either late or it is not. Outcomes like this can be characterized by a **binary variable** taking the value one for late payments and zero for those that are on time. Such variables are also called **indicator variables**, or **dummy variables**.

We summarize the probabilities of possible outcomes using a **probability density function** (***pdf***). The ***pdf*** for a discrete random variable indicates the **probability** of each possible value occurring. For a discrete random variable X the value of the probability density function $f(x)$ is the probability that the random variable X takes the value x, $f(x) = P(X = x)$. Because $f(x)$ is a probability, it must be true that $0 \leq f(x) \leq 1$ and, if X takes n possible values $x_1, \ldots, x_n$, then the sum of their probabilities must be one

$$P(X = x_1) + P(X = x_2) + \cdots + P(X = x_n) = f(x_1) + f(x_2) + \cdots + f(x_n) = 1$$

The **cumulative distribution function** (***cdf***) is an alternative way to represent probabilities. The *cdf* of the random variable X, denoted by $F(x)$, gives the probability that X is less than or equal to a specific value x. That is,

$$F(x) = P(X \leq x) \tag{B.1}$$

Two key features of a probability distribution are its center (location) and width (dispersion). A measure of the center is the **mean**, or **expected value**; measures of dispersion are **variance**, and its square root—the **standard deviation**.

B.1.1 Expected Value of a Discrete Random Variable

The **mean** of a random variable is given by its **mathematical expectation**. If X is a discrete random variable taking the values $x_1, \ldots, x_n$ then the mathematical expectation, or **expected value**,

of X is

$$\mu_X = E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_n P(X = x_n) \tag{B.2a}$$

The expected value, or mean, of X is a weighted average of its values, the weights being the probabilities that the values occur. The mean is often symbolized by μ or μ_X. It is the average value of the random variable in all possible experimental outcomes from the underlying **experiment**. Because the probability that the discrete random variable X takes the value x is given by its *pdf* $f(x)$, $P(X = x) = f(x)$, the expected value in (B.2a) can be written equivalently as

$$\begin{aligned}\mu_X = E(X) &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n) \\ &= \sum_{i=1}^{n} x_i f(x_i) = \sum_x x f(x)\end{aligned} \tag{B.2b}$$

Functions of random variables are also random. Expected values are obtained using calculations similar to those in (B.2). If X is a discrete random variable and $g(X)$ is a function of it, then

$$E[g(X)] = \sum_x g(x) f(x) \tag{B.3}$$

Using (B.3) we can develop some frequently used rules. If a is a constant, then

$$E(aX) = aE(X) \tag{B.4}$$

Similarly, if a and b are constants, then we can show that

$$E(aX + b) = aE(X) + b \tag{B.5}$$

To see how this result is obtained, we apply the definition in (B.3) to the function $g(X) = aX + b$

$$\begin{aligned}E[g(X)] &= \sum g(x) f(x) = \sum (ax + b) f(x) = \sum [ax f(x) + b f(x)] \\ &= \sum [ax f(x)] + \sum [b f(x)] = a \sum x f(x) + b \sum f(x) \\ &= aE(X) + b\end{aligned}$$

In the final step we recognize $E(X)$ from its definition in (B.2), and use the fact that $\sum f(x) = 1$.

If $g_1(X)$, $g_2(X)$, …, $g_M(X)$ are functions of X, then

$$E[g_1(X) + g_2(X) + \cdots + g_M(X)] = E[g_1(X)] + E[g_2(X)] + \cdots + E[g_M(X)] \tag{B.6}$$

This rule extends to any number of functions. **The expected value of a sum is always the sum of the expected values**.

A similar rule does not work, in general, for nonlinear functions. That is, $E[g(X)] \neq g[E(X)]$. For example, $E(X^2) \neq [E(X)]^2$.

B.1.2 Variance of a Discrete Random Variable

The **variance** of a discrete random variable X is the expected value of

$$g(X) = [X - E(X)]^2$$

The variance of a random variable is important in characterizing the scale of measurement and the spread of the probability distribution. We give it the symbol σ^2, which is read "sigma squared," or σ_X^2. Algebraically, letting $E(X) = \mu_X$,

$$\text{var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2 \tag{B.7}$$

The variance of a random variable is the *average* squared difference between the random variable X and its mean value μ. The larger the variance of a random variable, the more "spread out" its values are. The square root of the variance is called the **standard deviation**; it is denoted by σ or σ_X. It measures the spread or dispersion of a distribution and has the advantage of being in the same units of measure as the random variable.

A useful property of variances is the following. Let a and b be constants; then

$$\text{var}(aX + b) = a^2\text{var}(X) \tag{B.8}$$

This result is proven in the Probability Primer, Section P.5.4.

Two other characteristics of a probability distribution are its **skewness** and **kurtosis**. These are defined as

$$skewness = \frac{E\left[\left(X - \mu_X\right)^3\right]}{\sigma_X^3} \tag{B.9}$$

and

$$kurtosis = \frac{E\left[\left(X - \mu_X\right)^4\right]}{\sigma_X^4} \tag{B.10}$$

Skewness measures the lack of symmetry of a distribution. If the distribution is symmetric, then its *skewness* $= 0$. Distributions with long tails to the left are negatively skewed, and *skewness* < 0. Distributions with long tails to the right are positively skewed, and *skewness* > 0. Kurtosis measures the "peakedness" of a distribution. A distribution with large kurtosis has more values concentrated near the mean and a relatively high central peak. A distribution that is relatively flat has a lower kurtosis. The benchmark value for kurtosis is 3, which is the kurtosis of the **normal distribution** that we discuss later in this appendix (Section B.3.5).

B.1.3 Joint, Marginal, and Conditional Distributions

If X and Y are discrete random variables, then the joint probability that $X = a$ and $Y = b$ is given by the joint *pdf* of X and Y, written as $f(x, y)$, and $P[X = a, Y = b] = f(a, b)$. The sum of the joint probabilities is one, $\sum_x \sum_y f(x, y) = 1$. Given a **joint probability density function**, we can obtain the probability distributions of individual random variables, which are also known as **marginal distributions**. If X and Y are two discrete random variables, then

$$f_X(x) = \sum_y f(x, y) \text{ for each value } X \text{ can take} \tag{B.11}$$

For discrete random variables, the probability that the random variable Y takes the value y *given* that $X = x$ is written $P(Y = y|X = x)$. This conditional probability is given by the **conditional *pdf*** $f(y|x)$:

$$f(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x, y)}{f_X(x)} \tag{B.12}$$

Two random variables are **statistically independent** if the conditional probability that $Y = y$ given that $X = x$, is the same as the unconditional probability that $Y = y$ for all x and y values. In this case, knowing the value of X does not alter the probability distribution of Y. If X and Y are independent random variables, then

$$P(Y = y|X = x) = P(Y = y) \tag{B.13}$$

Equivalently, if X and Y are independent, then the conditional *pdf* of Y given $X = x$ is the same as the unconditional, or marginal, *pdf* of Y alone,

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = f_Y(y) \tag{B.14}$$

The converse is also true, so that if (B.13) or (B.14) is true *for every possible pair of x and y values*, then X and Y are statistically independent.

Solving (B.14) for the joint *pdf*, we can also say that X and Y are statistically independent if their joint *pdf* factors into the product of their marginal *pdf*s

$$f(x, y) = f_X(x) f_Y(y) \tag{B.15}$$

If (B.15) is true for each and every pair of x and y values, then X and Y are statistically independent. This result extends to more than two random variables. If X, Y, and Z are statistically independent, then their joint probability density function can be factored and written as $f(x, y, z) = f_X(x) \cdot f_Y(y) \cdot f_Z(z)$.

B.1.4 Expectations Involving Several Random Variables

A rule similar to (B.3) exists for functions of several random variables. Let X and Y be discrete random variables with joint *pdf* $f(x, y)$. If $g(X, Y)$ is a function of X and Y, then

$$E\big[g(X, Y)\big] = \sum_x \sum_y g(x, y) f(x, y) \tag{B.16}$$

Using (B.16) we can show that

$$E(X + Y) = E(X) + E(Y) \tag{B.17}$$

This follows by using the definition (B.16) and letting $g(X, Y) = X + Y$. Then

$$\begin{aligned}
E(X + Y) &= \sum_x \sum_y g(x, y) f(x, y) && \text{[general definition]} \\
&= \sum_x \sum_y (x + y) f(x, y) && \text{[specific function]} \\
&= \sum_x \sum_y x f(x, y) + \sum_x \sum_y y f(x, y) && \text{[separate terms]} \\
&= \sum_x x \sum_y f(x, y) + \sum_y y \sum_x f(x, y) && \text{[factor constants from 2nd sum]} \\
&= \sum_x x f(x) + \sum_y y f(y) && \text{[recognize marginal } pdf\text{]} \\
&= E(X) + E(Y) && \text{[recognize expected values]}
\end{aligned}$$

To go from the fourth to the fifth line, we have used (B.11) to obtain the marginal distributions of X and Y, and the fact that the order of summation does not matter. Using the same logic, we can show that

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{B.18}$$

In general, $E[g(X, Y)] \neq g[E(X), E(Y)]$. For example, in general, $E(XY) \neq E(X)E(Y)$. If, however, X and Y are statistically independent, then using (B.16), we can also show that $E(XY) = E(X)E(Y)$. To see this, recall that if X and Y are independent, then their joint *pdf* factors into the product of the marginal *pdf*s, $f(x, y) = f(x)f(y)$. Letting $g(X, Y) = XY$, we have

$$\begin{aligned}
E(XY) = E\big[g(X, Y)\big] &= \sum_x \sum_y xy f(x, y) = \sum_x \sum_y xy f(x) f(y) \\
&= \sum_x x f(x) \sum_y y f(y) = E(X)E(Y)
\end{aligned}$$

This rule can be extended to more independent random variables.

B.1.5 Covariance and Correlation

One particular application of (B.16) is the derivation of the **covariance** between X and Y. Define a function that is the product of X minus its mean times Y minus its mean,

$$g(X, Y) = \big(X - \mu_X\big)\big(Y - \mu_Y\big) \tag{B.19}$$

The covariance is the expected value of (B.19)

$$\text{cov}(X, Y) = \sigma_{XY} = E\Big[\big(X - \mu_X\big)\big(Y - \mu_Y\big)\Big] = E(XY) - \mu_X\mu_Y \tag{B.20}$$

If the covariance σ_{XY} of the variables is positive, then when x values are greater than their mean, the y values also tend to be greater than their mean, and when x values are below their mean, then the y values also tend to be less than their mean. In this case the random variables X and Y are said to be **positively** or **directly associated**. If $\sigma_{XY} < 0$, then the association is negative, or inverse. If $\sigma_{XY} = 0$, then there is neither a positive nor a negative relationship.

Interpreting the actual value of σ_{XY} is difficult, because X and Y may have different units of measurement. Scaling the covariance by the standard deviations of the variables eliminates the units of measurement, and defines the **correlation** between X and Y:

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{B.21}$$

As with the covariance, the correlation ρ between two random variables measures the degree of *linear* association between them. However, unlike the covariance, the correlation must lie between –1 and 1. The correlation between X and Y is 1 if there is a perfect positive linear relationship between X and Y and –1 if there is a perfect negative, or inverse, association between X and Y. If there is no *linear* association between X and Y, then $\text{cov}(X, Y) = 0$ and $\rho = 0$. For other values of correlation, the magnitude of the absolute value $|\rho|$ indicates the "strength" of the linear association between the values of the random variables.

If X and Y are independent random variables, then the covariance and correlation between them are zero. The converse of this relationship is *not* true. Independent random variables X and Y have zero covariance, indicating that there is no linear association between them. However, just because the covariance or correlation between two random variables is zero *does not* mean that they are necessarily independent. There may be more complicated nonlinear associations such as $X^2 + Y^2 = 1$.

In (B.17) we found the expected value of a sum of random variables. There are similar rules for variances. If a and b are constants, then

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\,\text{cov}(X, Y) \tag{B.22}$$

To see this, it is convenient to define a new discrete random variable $Z = aX + bY$. This random variable has expected value

$$\mu_Z = E(Z) = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

The variance of Z is

$$\begin{aligned}
\text{var}(Z) &= E\Big[\big(Z - \mu_Z\big)^2\Big] \\
&= E\Big\{\Big[(aX + bY) - \big(a\mu_X + b\mu_Y\big)\Big]^2\Big\} && \text{[substitute } Z\text{]} \\
&= E\Big\{\Big[\big(aX - a\mu_X\big) + \big(bY - b\mu_Y\big)\Big]^2\Big\} && \text{[combine like terms]}
\end{aligned}$$

$$= E\left\{\left[a(X-\mu_X)+b(Y-\mu_Y)\right]^2\right\} \quad \text{[factor]}$$

$$= E\left[a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y)\right] \quad \text{[expand]}$$

$$= E\left[a^2(X-\mu_X)^2\right]+E\left[b^2(Y-\mu_Y)^2\right]+E\left[2ab(X-\mu_X)(Y-\mu_Y)\right] \quad \text{[group terms]}$$

$$= a^2\text{var}(X)+b^2\text{var}(Y)+2ab\,\text{cov}(X,Y) \quad \text{[factor and recognize]}$$

These rules extend to more random variables. For example, if X, Y, and Z are random variables, then

$$\text{var}(aX+bY+cZ) = a^2\text{var}(X)+b^2\text{var}(Y)+c^2\text{var}(Z)+2ab\,\text{cov}(X,Y) + 2bc\,\text{cov}(Y,Z)+2ac\,\text{cov}(X,Z) \tag{B.23}$$

B.1.6 Conditional Expectations

If X and Y are two random variables with joint probability distribution $f(x, y)$, then the conditional probability distribution of Y given X is $f(y|x)$. We can use this conditional *pdf* to compute the **conditional mean** of Y given a value of X. That is, we can obtain the expected value of Y given that $X = x$. The conditional expectation $E(Y|X = x)$ is the average (or mean) value of Y given that X takes the value x. In the discrete case, it is defined to be

$$E(Y|X=x) = \sum_y yP(Y=y|X=x) = \sum_y yf(y|x) \tag{B.24}$$

Similarly, we can define the **conditional variance** of Y given X. This is the variance of the conditional distribution of Y given X. In the discrete case, it is

$$\text{var}(Y|X=x) = \sum_y \left[y-E(Y|X=x)\right]^2 f(y|x) \tag{B.25}$$

B.1.7 Iterated Expectations

The **law of iterated expectations** says that the expected value of Y is equal to the expected value of the conditional expectation of Y given X. That is,

$$E(Y) = E_X\left[E(Y|X)\right] \tag{B.26}$$

In Probability Primer Section P.6.3, we provide a numerical example of the Law of Iterated Expectations, and give the proof.

B.1.8 Variance Decomposition

Just as we can break up the expected value using the Law of Iterated Expectations, we can decompose the variance of a random variable into two parts.

$$\text{Variance Decomposition: } \text{var}(Y) = \text{var}_X\left[E(Y|X)\right] + E_X\left[\text{var}(Y|X)\right] \tag{B.27}$$

This result says that the variance of the random variable Y equals the sum of the variance of the conditional mean of Y given X and the mean of the conditional variance of Y given X. We discuss the **variance decomposition** for discrete random variables in Section P.6.4 of the Probability Primer. Here we provide the proof and a numerical example.

Proof of the Variance Decomposition We use the relationship between the marginal, conditional, and joint *pdf*s to prove the variance decomposition for discrete random variables. First, write out $\text{var}(Y)$ in an expanded form.

$$
\begin{aligned}
\text{var}(Y) &= \sum_y (y-\mu_y)^2 f(y) \\
&= \sum_y (y-\mu_y)^2 \left\{ \sum_x f(x,y) \right\} && \text{[replace marginal density]} \\
&= \sum_y (y-\mu_y)^2 \left\{ \sum_x f(y|x)\, f(x) \right\} && \text{[replace joint density]} \\
&= \sum_x \sum_y (y-\mu_y)^2 f(y|x)\, f(x) && \text{[change order of summation]} \\
&= \sum_x \sum_y \big(y - E(Y|x) + E(Y|x) - \mu_y\big)^2 f(y|x) f(x) && \text{[subtract and add conditional mean]} \\
&= \sum_x \sum_y \Big(\big[y - E(Y|x)\big] + \big[E(Y|x) - \mu_y\big] \Big)^2 f(y|x) f(x) && \text{[group terms, then square and expand]} \\
&= \sum_x \sum_y \Big\{ \big(y - E(Y|x)\big)^2 + \big(E(Y|x) - \mu_y\big)^2 + 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) \Big\} f(y|x) f(x) \\
&= \sum_x \sum_y \big(y - E(Y|x)\big)^2 f(y|x) f(x) && \textbf{[Term 1]} \\
&\quad + \sum_x \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) f(x) && \textbf{[Term 2]} \\
&\quad + \sum_x \sum_y 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) f(x) && \textbf{[Term 3]}
\end{aligned}
$$

Examine the three terms separately.

Term 3:

$$
\begin{aligned}
\textbf{Term 3} &= \sum_x \sum_y 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) f(x) \\
&= 2\sum_x \left\{ \sum_y \big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) \right\} f(x) && \text{[group inner sum]} \\
&= 2\sum_x \left\{ \big(E(Y|x) - \mu_y\big) \left[\sum_y \big(y - E(Y|x)\big) f(y|x) \right] \right\} f(x) && \text{[factor out constant]} \\
&= 2\sum_x \Big\{ \big(E(Y|x) - \mu_y\big)[0] \Big\} f(x) \\
&= 0
\end{aligned}
$$

In the third line above we recognize that in the summation over the values of y the expression $\big(E(Y|x) - \mu_y\big)$ does not vary, so that it can be factored out. The remaining term in the square brackets is zero because

$$
\begin{aligned}
&\sum_y \big(y - E(Y|x)\big) f(y|x) \\
&\quad = \sum_y y f(y|x) - E(Y|x) \sum_y f(y|x) && \big[\text{factor out the constant } E(Y|X)\big] \\
&\quad = E(Y|x) - E(Y|x) = 0 && \left[\text{definition of conditional expectation \& } \sum_y f(y|x) = 1\right]
\end{aligned}
$$

Term 2:

$$\begin{aligned}
\textbf{Term 2} &= \sum_x \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) f(x) \\
&= \sum_x \left\{ \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) \right\} f(x) \\
&= \sum_x \left\{ \big(E(Y|x) - \mu_y\big)^2 \sum_y f(y|x) \right\} f(x) && \left[\text{factor out } \big(E(Y|x) - \mu_y\big)^2\right] \\
&= \sum_x \left\{ \big(E(Y|x) - \mu_y\big)^2 \right\} f(x) && \left[\sum_y f(y|x) = \sum_y P\big(Y = y|X = x\big) = 1\right] \\
&= \sum_x \big(E(Y|x) - \mu_y\big)^2 f(x) \\
&= \text{var}_X\big[E(Y|X)\big]
\end{aligned}$$

In the final step, we label **Term 2** as $\text{var}_X[E(Y|X)] = \sum_x \big(E(Y|x) - \mu_y\big)^2 f(x)$. The intuition behind the terminology is discussed in Section P.6.3. The key point is that $E(Y|X)$ varies as the value of X varies. One way to recognize this is to say $E(Y|X) = g(X)$. Using first principles $\text{var}\big[g(X)\big] = E\big\{g(X) - E\big[g(X)\big]\big\}^2$. Also $E_X\big[g(X)\big] = E_X[E(Y|X)] = E(Y) = \mu_y$ using the law of iterated expectations. Then

$$\text{var}_X\big[g(X)\big] = E_X\left\{\big[g(X) - \mu_y\big]^2\right\} = E_X\left\{\big[E(Y|X) - \mu_y\big]^2\right\} = \sum_x \big[E(Y|x) - \mu_y\big]^2 f(x)$$

Term 1:

$$\begin{aligned}
\textbf{Term 1} &= \sum_x \sum_y \big(y - E(Y|x)\big)^2 f(y|x) f(x) \\
&= \sum_x \left\{ \sum_y \big(y - E(Y|x)\big)^2 f(y|x) \right\} f(x) \\
&= \sum_x \text{var}(Y|x) f(x) \\
&= E_X\big[\text{var}(Y|X)\big]
\end{aligned}$$

Term 1 is the expectation of the conditional variance of Y given X. A key point here, as in **Term 2**, is that the conditional variance of Y given X is a function of X.

EXAMPLE B.1 | Variance Decomposition: Numerical Example

The calculations illustrating the variance decomposition are somewhat involved. We have broken it up into parts to simplify the logic.

Variance of *Y*

For the population in Table P.1, given in the Probability Primer, the unconditional variance of Y is $\text{var}(Y) = E\big(Y^2\big) - \mu_Y^2$. We have shown that $E(Y) = \mu_Y = 2/5$. Also,

$$E\big(Y^2\big) = \sum_y y^2 f_Y(y) = 0^2 \times (6/10) + 1^2 \times (4/10) = 2/5$$

Then $\text{var}(Y) = E\big(Y^2\big) - \mu_Y^2 = 2/5 - (2/5)^2 = 6/25 = 0.24$.

Variance of the Conditional Expectation of *Y* Given *X*

The first component of the variance decomposition is $\text{var}_X[E(Y|X)]$. As we have noted earlier, $E(Y|X) = g(X)$ is a function of X. We computed these values to be $E(Y|X = 1) = 1$, $E(Y|X = 2) = 1/2$, $E(Y|X = 3) = 1/3$, and $E(Y|X = 4) = 1/4$. What is the variance of these terms, treating X as random? The variance of a function of X, $g(X)$, is

$$\text{var}_X\big[g(X)\big] = \sum_x \Big\{ g(x) - E_X\big[g(x)\big] \Big\}^2 f_X(x)$$

Using the law of iterated expectations

$$E_X\big[g(x)\big] = E_X[E(Y|X=x)] = E(Y).$$

The calculation we need is

$$\begin{aligned}\text{var}_X\big[E(Y|X)\big] &= \sum_x \big[E(Y|X=x) - \mu_Y\big]^2 f_X(x) \\ &= \left[\sum_x E(Y|X=x)^2 f_X(x)\right] - \mu_Y^2\end{aligned}$$

Now

$$\begin{aligned}&\sum_x E(Y|X=x)^2 f_X(x) \\ &= E(Y|X=1)^2 f_X(1) + E(Y|X=2)^2 f_X(2) \\ &\quad + E(Y|X=3)^2 f_X(3) + E(Y|X=4)^2 f_X(4) \\ &= 1^2\left(\frac{1}{10}\right) + \left(\frac{1}{2}\right)^2\left(\frac{2}{10}\right) + \left(\frac{1}{3}\right)^2\left(\frac{3}{10}\right) + \left(\frac{1}{4}\right)^2\left(\frac{4}{10}\right) \\ &= \frac{5}{24}\end{aligned}$$

Then,

$$\begin{aligned}\text{var}_X\big[E(Y|X)\big] &= \left[\sum_x E(Y|X=x)^2 f_X(x)\right] - \mu_Y^2 = \frac{5}{24} - \left(\frac{2}{5}\right)^2 \\ &= \frac{29}{600} = 0.048333\ldots\end{aligned}$$

That is, $E(Y|X)$ exhibits variation as X changes and has variance 0.0483.

Expectation of the Conditional Variance of *Y* Given *X*

The second component of the variance decomposition is $E_X[\text{var}(Y|X)]$. The conditional variance $\text{var}(Y|X=x)$ varies randomly as X varies, if we treat X as random, so that finding its expected value makes sense. For the population in Table P.1, we have already computed the conditional means $E(Y|X=x)$ for each x. The conditional variances are $\text{var}(Y|X=x) = E\big(Y^2|X=x\big) - [E(Y|X=x)]^2$ so we need the terms $E\big(Y^2|X=x\big)$ for each value of X. These are

$$\begin{aligned}&E\big(Y^2|X=1\big) = 1, \quad E\big(Y^2|X=2\big) = 1/2, \\ &E\big(Y^2|X=3\big) = 1/3, \quad E\big(Y^2|X=4\big) = 1/4\end{aligned}$$

Then

$$\begin{aligned}\text{var}(Y|X=1) &= E\big(Y^2|X=1\big) - \big[E(Y|X=1)\big]^2 \\ &= 1 - 1^2 = 0 \\ \text{var}(Y|X=2) &= E\big(Y^2|X=2\big) - \big[E(Y|X=2)\big]^2 \\ &= 1/2 - (1/2)^2 = 1/4 \\ \text{var}(Y|X=3) &= E\big(Y^2|X=3\big) - \big[E(Y|X=3)\big]^2 \\ &= 1/3 - (1/3)^2 = 2/9 \\ \text{var}(Y|X=4) &= E\big(Y^2|X=4\big) - \big[E(Y|X=4)\big]^2 \\ &= 1/4 - (1/4)^2 = 3/16\end{aligned}$$

The expected value of the conditional variance is

$$\begin{aligned}E_X\big[\text{var}(Y|X)\big] &= \sum_x \text{var}(Y|X=x) f_X(x) \\ &= 0(1/10) + (1/4)(2/10) \\ &\quad + (2/9)(3/10) + (3/16)(4/10) \\ &= 23/120 = 0.191666\ldots\end{aligned}$$

The interpretation of this expectation is that if we repeatedly drew a random member from the population in Table P.1, and for each value computed the conditional variance $\text{var}(Y|X=x)$, the average of the conditional variance in many trials would approach 0.19167.

Variance of *Y* Decomposed

We have shown that for the population in Table P.1 $\text{var}_X[E(Y|X)] = 29/600$ and $E_X[\text{var}(Y|X)] = 23/120$. The variance decomposed is

$$\begin{aligned}\text{var}(Y) &= \text{var}_X\big[E\big(Y|X\big)\big] + E_X\big[\text{var}(Y|X)\big] \\ &= \frac{29}{600} + \frac{23}{120} = \frac{144}{600} = \frac{6}{25} = 0.24\end{aligned}$$

This is the same value for $\text{var}(Y)$ that we derived in the first step above.

B.1.9 Covariance Decomposition

Recall that the covariance between two random variables Y and X is $\text{cov}(X,Y) = E\big[\big(X-\mu_X\big)\big(Y-\mu_Y\big)\big]$. For discrete random variables the expectation is

$$\text{cov}(X,Y) = \sum_x \sum_y \big(x-\mu_X\big)\big(y-\mu_Y\big) f(x,y)$$

By using the relationships between marginal, conditional, and joint *pdf*s we can show

$$\text{cov}(X,Y) = \sum_x \big(x-\mu_X\big) E(Y|X=x) f(x)$$

Recall that $E(Y|X) = g(X)$ is a function of X. The covariance between X and Y can be calculated as the expected value of X, minus its mean, times a function of X,

$$\text{cov}(X, Y) = E_X\Big[(X - \mu_X)E(Y|X)\Big] \tag{B.28}$$

A numerical example of this **covariance decomposition** is given in the Probability Primer Section P.6.5.

Proof of the Covariance Decomposition

$$\begin{aligned}\text{cov}(X, Y) &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ &= \sum_x \sum_y (x - \mu_X) y f(x, y) - \mu_Y \sum_x \sum_y (x - \mu_X) f(x, y)\end{aligned}$$

In this expression, the second term is zero, because

$$\begin{aligned}\sum_x \sum_y (x - \mu_X) f(x, y) &= \sum_x (x - \mu_X) \sum_y f(x, y) && \Big[\text{factor out } (x - \mu_X)\Big] \\ &= \sum_x (x - \mu_X) f(x) && \Big[\sum_y f(x, y) = f(x)\Big] \\ &= \sum_x x f(x) - \mu_X \sum_x f(x) \\ &= \mu_X - \mu_X = 0 && \Big[\sum_x f(x) = 1\Big]\end{aligned}$$

Then

$$\begin{aligned}\text{cov}(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) &= \sum_x \sum_y (x - \mu_X) y f(x, y) \\ &= \sum_x (x - \mu_X) \Big\{ \sum_y y f(y|x) \Big\} f(x) \\ &= \sum_x (x - \mu_X) E(Y|X = x) f(x)\end{aligned}$$

B.2 Working with Continuous Random Variables

Continuous random variables can take any value in at least one interval. In economics, variables like income and market prices are treated as continuous random variables. In Figure P.2 of the Probability Primer, we depict the probability density function for a continuous random variable that ranges between zero and infinity, or $x > 0$. Because continuous random variables can take uncountably many values, the probability that any single value occurs in a random experiment is zero. For example, $P(X = 100) = 0$ or $P(X = 200) = 0$. Probability statements for continuous random variables are meaningful when we ask about outcomes within intervals, or ranges. We can ask, "What is the probability that X takes a value between 100 and 200?" These ideas were introduced in Sections P.1 and P.2 of the Probability Primer. There we noted that probabilities like these are areas under a curve that is the probability density function. It would be a good time to review those sections now if the concepts are not fresh in your minds. What we did not discuss in the Probability Primer was how exactly such probabilities are calculated. We delayed that discussion until now, because tools from integral calculus are required.

In this section, we discuss how to work with continuous random variables. The interpretation of probabilities, expected values, and variances carries over from what you learned about discrete random variables. What changes is the algebra—summation signs turn into integrals,

and this takes a little getting used to. If you have not done so, review the discussion of integrals in Appendix A.4.

B.2.1 Probability Calculations

If X is a continuous random variable with probability density function $f(x)$, then $f(x)$ must obey certain properties:

$$f(x) \geq 0 \tag{B.29}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{B.30}$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx \tag{B.31a}$$

Property (B.29) states that the *pdf* cannot take negative values. Property (B.30) states that the total area under the *pdf*, which is the probability that X falls between $-\infty$ and ∞, is one. Property (B.31a) states that the probability that X falls in the interval $[a, b]$ is the area under the curve $f(x)$ between those values. Because a single point has probability zero, it is also true that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_{a}^{b} f(x)dx \tag{B.31b}$$

The **cumulative distribution function**, *cdf*, for a continuous random variable is $F(x) = P(X \leq x)$. Using the *cdf* we can compute

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx = F(a) \tag{B.32a}$$

The *cdf* is obtained by integrating the *pdf*. The integral is an "antiderivative," so that we can obtain the *pdf* $f(x)$ by differentiating the *cdf* $F(x)$. That is,

$$f(x) = \frac{dF(x)}{dx} = F'(x) \tag{B.32b}$$

The concept of a *cdf* is useful in many ways, including working with computer software, which includes the *cdf*s of many random variables so that probabilities can be easily computed.

EXAMPLE B.2 | Probability Calculation Using Geometry

Let X be a continuous random variable with *pdf* $f(x) = 2(1 - x)$ for $0 \leq x \leq 1$. This *pdf* is depicted in Figure B.1.

Property (B.29) holds for x in the interval [0, 1]. Furthermore, property (B.30) holds because

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} 2(1-x)dx = \int_{0}^{1} 2dx - \int_{0}^{1} 2xdx$$

$$= 2x\Big|_0^1 - x^2\Big|_0^1 = 2 - 1 = 1$$

Using Figure B.1, we can compute $P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = \frac{1}{2}$ using geometry. Using integration, we come to the same conclusion:

$$P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = \int_{1/4}^{3/4} f(x)dx = \int_{1/4}^{3/4} 2(1-x)dx$$

$$= \int_{1/4}^{3/4} 2dx - \int_{1/4}^{3/4} 2xdx = 2x\Big|_{1/4}^{3/4} - x^2\Big|_{1/4}^{3/4}$$

$$= 1 - \left(\frac{9}{6} - \frac{1}{16}\right) = \frac{1}{2}$$

The cumulative distribution function is $F(x) = 2x - x^2$ for x in the interval [0, 1], so the probability can also be computed as

$$P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = F\left(\frac{3}{4}\right) - F\left(\frac{1}{4}\right)$$

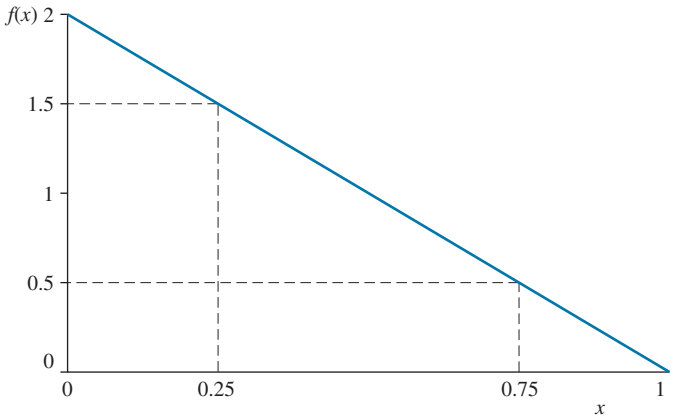


FIGURE B.1 **Probability density function $f(x) = 2(1-x)$.**

EXAMPLE B.3 | Probability Calculation Using Integration

Let X be a continuous random variable with *pdf* $f(x) = 3x^2$ for x in the interval [0,1]. Properties (B.29) and (B.30) hold. Because the *pdf* is a quadratic, we cannot use simple geometry to compute $P\left(\frac{1}{4} \le X \le \frac{3}{4}\right)$. We must use integration, obtaining

$$P\left(\frac{1}{4} \le X \le \frac{3}{4}\right) = \int_{1/4}^{3/4} f(x)\,dx = \int_{1/4}^{3/4} 3x^2 dx = x^3\Big|_{1/4}^{3/4}$$
$$= \frac{9}{64} - \frac{1}{64} = \frac{1}{8}$$

B.2.2 Properties of Continuous Random Variables

If X is a continuous random variable with probability density function $f(x)$, then its expected value is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{B.33}$$

Compare this to the expected value of a discrete random variable in (B.2). An integral has replaced the summation. The interpretation of $E(X)$ is exactly the same as in the discrete case. It is the average value of X that occurs in all possible samples from an underlying experiment.

EXAMPLE B.4 | Expected Value of a Continuous Random Variable

The expected value of the random variable in Example B.2 is

$$\int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x \cdot 2(1-x)\,dx = \int_0^1 \left(2x - 2x^2\right) dx = x^2\Big|_0^1 - \tfrac{2}{3}x^3\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

The variance of a random variable X is defined as $\sigma_X^2 = E\left[\left(X - \mu_X\right)^2\right]$. This definition holds for discrete and continuous random variables. In order to compute the variance we use the analog to the rule in (B.3) for continuous random variables,

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{B.34}$$

Then, letting $g(x) = (X - \mu_X)^2$, we have

$$\begin{aligned}\sigma_X^2 &= E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty}(x-\mu_X)^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty}\left(x^2 + \mu_X^2 - 2x\mu_X\right) f(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx + \mu_X^2\int_{-\infty}^{\infty} f(x)\,dx - 2\mu_X\int_{-\infty}^{\infty} x f(x)\,dx \\ &= E(X^2) + \mu_X^2 - 2\mu_X^2 \\ &= E(X^2) - \mu_X^2 \end{aligned} \tag{B.35}$$

To go from the third line to the fourth line, we use property (B.30) and the definition of expected value (B.33). The end result is that $\sigma_X^2 = E\left[(X-\mu_X)^2\right] = E(X^2) - \mu_X^2$ as in the discrete case.

EXAMPLE B.5 | Variance of a Continuous Random Variable

To obtain the variance of the random variable described in Example B.2, we first find

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 \cdot 2(1-x)dx = \int_0^1 (2x^2 - 2x^3)dx$$
$$= \tfrac{2}{3}x^3\Big|_0^1 - \tfrac{2}{4}x^4\Big|_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$$

Then,

$$\text{var}(X) = \sigma_X^2 = E(X^2) - \mu_X^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18}$$

B.2.3 Joint, Marginal, and Conditional Probability Distributions

To make simultaneous probability statements about more than one continuous random variable, we need the **joint probability density function** of the random variables. For example, consider the two continuous random variables U = unemployment and P = inflation rate. Suppose that their joint *pdf* is as depicted in Figure B.2.

The joint *pdf* is a surface and probabilities are volumes under the surface. If the two random variables are nonnegative, then we might ask, "What is the probability that inflation is less than 5% and at the same time unemployment is less than 6%?" That is, what is $P(U \le 6, P \le 5)$? Geometrically the answer is that this is the volume under the surface above the rectangle (in the base of the figure) defining the event. Just as an integral is used to obtain the area under a curve, a double integral is used to obtain volumes like that shown in Figure B.2. Given the joint *pdf* $f(u, p)$ we can compute the probability as

$$P(U \le 6, P \le 5) = \int_{u=0}^{6}\int_{p=0}^{5} f(u,p)\,dp\,du$$

If we know the joint *pdf*, can we obtain the marginal *pdf* of one of the random variables? If so, we can answer questions like "What is the probability that unemployment will be between 2% and 5%?" Analogous to (B.11) for discrete random variables, we integrate out the unwanted random variable. That is, the **marginal probability density function** for U is

$$f(u) = \int_{-\infty}^{\infty} f(u,p)\,dp \tag{B.36}$$

Then, for example, $P(2 \le U \le 5) = \int_2^5 f(u)\,du$.

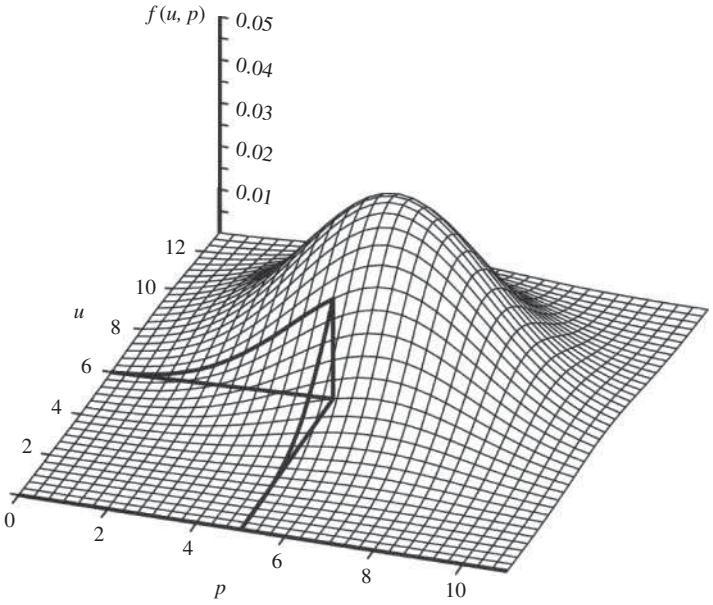


FIGURE B.2 **A joint probability density function.**

We might ask "What is the probability that unemployment will be between 2% and 5% **if** we can use monetary policy to keep the inflation rate at 2%?" This is a question about a **conditional probability**. Given that $P = 2$, what is the probability that $2 \leq U \leq 5$? Or in terms of conditioning notation, what is $P(2 \leq U \leq 5|P = 2)$? To answer such questions for continuous random variables, we need the **conditional probability density function** $f(u|p)$, which is given by

$$f(u|p) = \frac{f(u,p)}{f(p)} \tag{B.37}$$

Unlike the result (B.12) for discrete random variables, we do not obtain the probability from this division, but rather a density function that can be used for probability calculations. Not only can we obtain conditional probabilities using $f(u|p)$, but we can also obtain the **conditional expectation**, or **conditional mean**,

$$E(U|P = p) = \int_{-\infty}^{\infty} u f(u|p)\, du \tag{B.38}$$

Similarly, the **conditional variance** is

$$\text{var}(U|P = p) = \int_{-\infty}^{\infty} \left[u - E(U|P = p)\right]^2 f(u|p)\, du \tag{B.39}$$

The importance of questions involving unemployment and inflation are of great social importance. Economists and econometricians work on these problems, and you will glimpse the issues a few times throughout this book. But it is difficult. So we illustrate the above concepts with a simpler example.

EXAMPLE B.6 | Computing a Joint Probability

Let X and Y be continuous random variables with joint *pdf* $f(x, y) = x + y$ for x in [0, 1] and y in [0, 1]. You might test your geometric skills by creating a three-dimensional graph of this joint density function. Is it a valid density function? It satisfies the more general version of property (B.29), because $f(x, y) \geq 0$ for all points $x \in [0, 1]$ and $y \in [0, 1]$. Also the total amount of probability, the volume under the surface, is

$$\begin{aligned}
\int_{y=0}^{1}\int_{x=0}^{1} f(x,y)dx\,dy &= \int_{y=0}^{1}\int_{x=0}^{1}(x+y)dx\,dy \\
&= \int_{y=0}^{1}\int_{x=0}^{1} xdx\,dy + \int_{y=0}^{1}\int_{x=0}^{1} ydx\,dy \\
&= \int_{y=0}^{1}\left[\int_{x=0}^{1} xdx\right]dy + \int_{x=0}^{1}\left[\int_{y=0}^{1} ydy\right]dx \\
&= \int_{y=0}^{1}\left[\tfrac{1}{2}x^2\Big|_0^1\right]dy + \int_{x=0}^{1}\left[y^2\Big|_0^1\right]dx \\
&= \int_{y=0}^{1}\tfrac{1}{2}dy + \int_{x=0}^{1}\tfrac{1}{2}dx = \frac{1}{2}+\frac{1}{2} = 1
\end{aligned}$$

In the third line, we have used a property of multiple integrals. In the Probability Primer, Section P.4, the rule "Sum 9" states that the order of multiple summations does not matter. Similarly, as long as the limits of integration for one variable do not depend on the value of the other, the order of integration does not matter when we have multiple integrals. However, we must keep the integral symbol with its lower and upper limits paired with the variable of integration, indicated by dx or dy. In the first term in the third line above, we have isolated the integral involving x inside the integral involving y. Multiple integrals are evaluated by working from the "inside out." Solve the inside integral with respect to x, and then solve the outer integral with respect to y.

EXAMPLE B.7 | Another Joint Probability Calculation

For further practice with double integrals find the probability that X is between zero and 1/2 while Y is between 1/4 and 3/4 for the joint *pdf* in Example B.6. This is a joint probability and is computed as follows:

$$\begin{aligned}
&P\left(0 \leq X \leq \tfrac{1}{2}, \tfrac{1}{4} \leq Y \leq \tfrac{3}{4}\right) \\
&= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\int_{x=0}^{\frac{1}{2}} f(x,y)dx\,dy \\
&= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\int_{x=0}^{\frac{1}{2}}(x+y)dx\,dy \\
&= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\left[\int_{x=0}^{\frac{1}{2}} xdx\right]dy + \int_{y=\frac{1}{4}}^{\frac{3}{4}} y\left[\int_{x=0}^{\frac{1}{2}} dx\right]dy \\
&= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\left[\tfrac{1}{2}x^2\Big|_0^{1/2}\right]dy + \int_{y=\frac{1}{4}}^{\frac{3}{4}} y\left[x\Big|_0^{1/2}\right]dy \\
&= \tfrac{1}{8}\int_{y=\frac{1}{4}}^{\frac{3}{4}} dy + \tfrac{1}{2}\int_{y=\frac{1}{4}}^{\frac{3}{4}} ydy \\
&= \tfrac{1}{8}\left(y\Big|_{1/4}^{3/4}\right) + \tfrac{1}{2}\left(\tfrac{1}{2}y^2\Big|_{1/4}^{3/4}\right) \\
&= \frac{1}{8}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{4} = \frac{3}{16}
\end{aligned}$$

In the third step of this example, we did not change the order of integration in the second term. This illustrates another feature of working with multiple integrals. When carrying out the "inside" integration with respect to x the value of y is fixed, and because it is fixed it can be factored out, leaving a simpler inside integral.

EXAMPLE B.8 | Finding and Using a Marginal *pdf*

The marginal *pdf* of X, for $x \in [0, 1]$, is

$$f(x) = \int_{y=0}^{1} f(x,y)\,dy = \int_{y=0}^{1} (x+y)\,dy$$
$$= \int_{y=0}^{1} x\,dy + \int_{y=0}^{1} y\,dy = x{\cdot}y\Big|_0^1 + \tfrac{1}{2}y^2\Big|_0^1 = x + \frac{1}{2}$$

Technically we should also say that $f(x) = 0$ for $x \notin [0, 1]$, but we will generally not explicitly include this extra information. Using similar steps the marginal *pdf* of Y is $f(y) = y + 1/2$ for values of y in the [0, 1] interval. The marginal *pdf* for X can be used to compute probabilities that X falls in intervals in the domain of X, $x \in [0, 1]$. For example,

$$P\left(\tfrac{1}{2} < X < \tfrac{3}{4}\right) = \int_{1/2}^{3/4} \left(x + \tfrac{1}{2}\right) dx = \int_{1/2}^{3/4} x\,dx + \tfrac{1}{2}\int_{1/2}^{3/4} dx$$
$$= \tfrac{1}{2}x^2\Big|_{1/2}^{3/4} + \tfrac{1}{2}x\Big|_{1/2}^{3/4}$$
$$= \frac{1}{2}\left(\frac{9}{16} - \frac{1}{4}\right) + \frac{1}{2}\left(\frac{3}{4} - \frac{1}{2}\right)$$
$$= \frac{1}{2} \times \frac{5}{16} + \frac{1}{2} \times \frac{1}{4} = \frac{9}{32}$$

Using the marginal *pdf* of X, we can find its expected value.

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_0^1 x\left(x + \tfrac{1}{2}\right) dx$$
$$= \int_0^1 x^2 dx + \int_0^1 \tfrac{1}{2}x dx$$
$$= \tfrac{1}{3}x^3\Big|_0^1 + \tfrac{1}{4}x^2\Big|_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

The limits of integration in the first line change from $(-\infty, \infty)$ to [0, 1], because for $x \notin [0, 1]$, $f(x) = 0$ and the area (probability) under $f(x) = 0$ is zero.

To find the variance of X, we first find

$$E\left(X^2\right) = \int_0^1 x^2 f(x)dx = \int_0^1 x^2\left(x + \tfrac{1}{2}\right)dx$$
$$= \int_0^1 x^3 dx + \int_0^1 \tfrac{1}{2}x^2 dx$$
$$= \tfrac{1}{4}x^4\Big|_0^1 + \tfrac{1}{6}x^3\Big|_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

Then

$$\sigma_X^2 = \text{var}(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

The conditional *pdf* of Y given that $X = x$ is $f(y|x) = f(x, y)/f(x)$.

EXAMPLE B.9 | Finding and Using a Conditional *pdf*

In Example B.6 the conditional *pdf* is

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{x+y}{x+\frac{1}{2}} \quad \text{for} \quad y \in [0, 1]$$

As a specific example,

$$f\left(y\middle|X = \tfrac{1}{3}\right) = \frac{y + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}} = \frac{1}{5}(6y + 2) \quad \text{for} \quad y \in [0, 1]$$

The conditional *pdf* can be used to compute probabilities that Y falls in a given interval. Also, we can compute the **conditional mean** of Y given that $X = 1/3$

$$\mu_{Y|X=1/3} = E\left(Y\middle|X = \tfrac{1}{3}\right) = \int_{y=0}^{1} y f\left(y\middle|X = \tfrac{1}{3}\right)dy$$
$$= \int_{y=0}^{1} y\tfrac{1}{5}(6y + 2)dy$$
$$= \int_{y=0}^{1} \tfrac{6}{5}y^2 dy + \int_{y=0}^{1} \tfrac{2}{5}y dy$$
$$= \tfrac{6}{5}\left(\tfrac{1}{3}y^3\Big|_0^1\right) + \tfrac{2}{5}\left(\tfrac{1}{2}y^2\Big|_0^1\right) = \frac{2}{5} + \frac{1}{5} = \frac{3}{5}$$

Note that the conditional expected value is not the same as the **unconditional** expected value $\mu_Y = E(Y) = \frac{7}{12}$. To calculate the **conditional variance**, we first calculate

$$E\left(Y^2 \left| X = \tfrac{1}{3}\right.\right) = \int_{y=0}^{1} y^2 f\left(y \left| X = \tfrac{1}{3}\right.\right) dy$$

$$= \int_{y=0}^{1} y^2 \tfrac{1}{5}(6y + 2) dy = \frac{13}{30}$$

The conditional variance is then

$$\text{var}\left(Y \left| X = \tfrac{1}{3}\right.\right) = E\left(Y^2 \left| X = \tfrac{1}{3}\right.\right) - \left[E\left(Y \left| X = \tfrac{1}{3}\right.\right)\right]^2$$

$$= \frac{11}{150} = 0.07333$$

The unconditional variance is $\sigma_Y^2 = \text{var}(Y) = \frac{11}{144} = 0.07639$.

The **correlation** between X and Y is

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

The covariance between X and Y can be calculated using $\text{cov}(X, Y) = E(XY) - \mu_X \mu_Y$.

EXAMPLE B.10 | Computing a Correlation

To compute the expected value of XY for Example B.6, we calculate the double integral

$$E(XY) = \int_{y=0}^{1} \int_{x=0}^{1} xy f(x, y) dx dy$$

$$= \int_{y=0}^{1} \int_{x=0}^{1} xy(x + y) dx dy$$

$$= \int_{y=0}^{1} \int_{x=0}^{1} x^2 y dx dy + \int_{y=0}^{1} \int_{x=0}^{1} xy^2 dx dy$$

$$= \int_{y=0}^{1} y \left[\int_{x=0}^{1} x^2 dx\right] dy + \int_{y=0}^{1} y^2 \left[\int_{x=0}^{1} x dx\right] dy$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Then

$$\text{cov}(X, Y) = E(XY) - \mu_X \mu_Y = \frac{1}{3} - \left(\frac{7}{12}\right)\left(\frac{7}{12}\right) = \frac{-1}{144}$$

Finally, the correlation between X and Y is

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-1/144}{\sqrt{11/144}\sqrt{11/144}} = \frac{-1}{11} = -0.09091$$

B.2.4 Using Iterated Expectations with Continuous Random Variables

A useful result, proved in Section B.1.7 for the discrete case, is the **law of iterated expectations**. If X and Y are continuous random variables with joint *pdf* $f(x,y)$, then the expected value of Y can be calculated as

$$E(Y) = E_X\left[E(Y|X)\right]$$

This is the same result as in (B.26) for the discrete case. The exact meaning of this expression is best understood by first deriving it and then carrying through an illustration. To establish that this result is true, we proceed as follows:

$$
\begin{aligned}
E(Y) &= \int_{y=-\infty}^{\infty} y f(y) dy \\
&= \int_{y=-\infty}^{\infty} y \left[\int_{x=-\infty}^{\infty} f(x,y) dx \right] dy && \text{[replacing marginal } pdf\text{]} \\
&= \int_{y} \int_{x} y f(x,y) dx\, dy && \text{[simplifying integral]} \\
&= \int_{y} \int_{x} y \big[f(y|x) f(x) \big] dx\, dy && \text{[replace joint } pdf\text{]} \\
&= \int_{x} \left[\int_{y} y f(y|x) dy \right] f(x) dx && \text{[reverse order of integration]} \\
&= \int_{x} \big[E(Y|X) \big] f(x) dx && \text{[recognize } E(Y|X)\text{]} \\
&= E_X \big[E(Y|X) \big] && \text{[recognize expectation wrt } X\text{]}
\end{aligned}
$$

In the last line of the expression, the notation $E_X[\cdot]$ means that we take the expectation of the term in brackets treating X as random. Note that we also replaced the $(-\infty, \infty)$ integral form with a simpler form in line three indicating "over all values" of the variable of integration.

EXAMPLE B.11 | Using Iterated Expectation

To better understand the **iterated expectation** expression, for Example B.6 find the **conditional expectation** of Y given that $X = x$, where the value x is not specified:

$$
\begin{aligned}
E(Y|X = x) &= \int_{y=0}^{1} y f(y|x)\, dy = \int_{y=0}^{1} y \left[\frac{x+y}{x+\frac{1}{2}} \right] dy \\
&= \frac{2+3x}{3(2x+1)}
\end{aligned}
$$

Note that the integration over the values of Y, treating x as given, leaves us with a function of x. If we now recognize that x can take any value and is thus random, we can find the expected value of the function

$$
g(X) = \frac{2+3X}{3(2X+1)}
$$

The law of iterated expectations says that if we take the expectation of $g(X)$, treating X as random, we should obtain $E(Y)$.

$$
\begin{aligned}
E\big[g(X)\big] &= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} f(x) dx \\
&= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} \left(x + \frac{1}{2} \right) dx \\
&= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} \frac{1}{2} (2x+1) dx = \int_{x=0}^{1} \tfrac{1}{6} (2+3x) dx \\
&= \int_{x=0}^{1} \tfrac{1}{3} dx + \int_{x=0}^{1} \tfrac{1}{2} x dx = \tfrac{1}{3} x \Big|_0^1 + \tfrac{1}{4} x^2 \Big|_0^1 \\
&= \frac{1}{3} + \frac{1}{4} = \frac{7}{12} = E(Y)
\end{aligned}
$$

There are several important implications of the law of iterated expectations. First, based on $E(Y) = E_X\big[E(Y|X)\big]$, we can see that if $E(Y|X) = 0$, then $E(Y) = E_X[E(Y|X)] = E_X(0) = 0$. If the conditional expectation of Y is zero, then the unconditional expectation of Y is also zero.

Second, if $E(Y|X) = E(Y)$, then $\text{cov}(X, Y) = 0$. To see this, first rewrite $E(XY)$ as

$$\begin{aligned} E(XY) &= \int_x \int_y xyf(x,y)dydx \\ &= \int_x \int_y xyf(y|x)f(x)dydx \\ &= \int_x x\left[\int_y yf(y|x)dy\right]f(x)dx \\ &= \int_x x\big[E(Y|X)\big]f(x)dx \end{aligned} \tag{B.40}$$

If $E(Y|X) = E(Y)$, then the last line of (B.40) becomes

$$\begin{aligned} E(XY) &= \int_x x\big[E(Y)\big]\, f(x)dx = E(Y)\int_x xf(x)dx \\ &= E(Y)E(X) = \mu_Y\mu_X \end{aligned}$$

The covariance between X and Y in this case is

$$\text{cov}(X, Y) = E(XY) - \mu_X\mu_Y = \mu_X\mu_Y - \mu_Y\mu_X = 0$$

An extremely important special case of these two results concerns the consequences of $E(Y|X) = 0$. We have already seen that $E(Y|X) = 0 \Rightarrow E(Y) = 0$. Now we can also see that if $E(Y|X) = E(Y) = 0$, then $\text{cov}(X, Y) = 0$.

B.2.5 Distributions of Functions of Random Variables

As we have noted several times, a function of a random variable is random itself. The question we address in this section is, "What is the probability density function of the new random variable?" For the case of a discrete random variable this problem is not too hard. For example, consider the discrete random variable X that can take the values 1, 2, 3, or 4 with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Let $Y = 2 + 3X = g(X)$. What is the *pdf* for Y? In this case it is clear. The probability that $Y = 5, 8, 11$, or 14 corresponds exactly to the probability that $X = 1, 2, 3$, or 4, respectively, as shown in Table B.1.

What makes this possible is that each value of y corresponds to a unique value of x, and each value of x corresponds to a unique value of y. Another way to say this is that the transformation from X to Y is "one-to-one." This type of relationship is ensured to hold when the function $g(X)$ relating Y to X is either strictly increasing or strictly decreasing. Such functions are said to be *strictly monotonic*. Our function $Y = 2 + 3X = g(X)$ is strictly (monotonically) increasing. This guarantees that if $x_2 > x_1$, then $y_2 = g(x_2) > y_1 = g(x_1)$. Note in particular that we are ruling out the possibility that $y_1 = y_2$.

Determining the distribution of $Y = g(X)$ in the continuous case is a bit more challenging. In the following example, we present the **change-of-variable** technique that applies when the function $g(X)$ is strictly increasing or decreasing.

TABLE B.1 **Change of Variable: Discrete Case**

x	$P(X = x) = P(Y = y)$	y
1	0.1	5
2	0.2	8
3	0.3	11
4	0.4	14

EXAMPLE B.12 | Change of Variable: Continuous Case

Let X be a continuous random variable with *pdf* $f(x) = 2x$ for $0 < x < 1$. Let $Y = g(X) = 2X$ be another random variable. We want to compute probabilities that Y falls in certain intervals. One solution is to compute probabilities for Y based on the probability of the corresponding event for X. For example,

$$P(0 < Y < 1) = P\left(0 < X < \frac{1}{2}\right) = \int_0^{1/2} 2x\,dx = x^2\Big|_0^{1/2} = \frac{1}{4}$$

Although this is reasonable and relatively simple in this case, it will not always be so. It is preferable to determine the *pdf* of Y, say $h(y)$, and use it to compute probabilities for Y. Since $X = Y/2$, we might be tempted to substitute this into the *pdf* $f(x)$ to obtain $h(y) = 2(y/2) = y$ for $0 < y < 2$. This substitution does not work, however, because

$$\int_{-\infty}^{\infty} h(y)\,dy = \int_0^2 y\,dy = \tfrac{1}{2}y^2\Big|_0^2 = 2$$

This violates property (B.30) for a probability density function. Furthermore, using $h(y)$ to compute the probability of Y falling in the interval (0, 1) produces 0.5, which we know is incorrect.

The problem is that we must adjust the height of $h(y)$ to account for the fact that Y can take values in the interval (0, 2) whereas X can take values only in (0, 1). In fact, a change in Y of one unit corresponds to a change in X of half a unit. If we adjust $h(y)$ by this factor, we have

$$h(y) = 2(y/2)\left(\frac{1}{2}\right) = y/2, \quad 0 < y < 2$$

Using this corrected *pdf*, property (B.30) is satisfied:

$$\int_{-\infty}^{\infty} h(y)\,dy = \int_0^2 \tfrac{1}{2}y\,dy = \tfrac{1}{4}y^2\Big|_0^2 = 1$$

Also, we obtain the correct probability that Y falls in the interval (0, 1):

$$P(0 < Y < 1) = \int_0^1 \tfrac{1}{2}y\,dy = \tfrac{1}{4}y^2\Big|_0^1 = \frac{1}{4}$$

Another perspective on the change-of-variable technique is obtained by examining the integral representation for the probability that Y falls in the interval (0, 1):

$$P(0 < Y < 1) = \int_0^1 h(y)\,dy$$

The integral representation of the equivalent X event, showing explicitly the lower and upper limits of the integral, is

$$P(0 < Y < 1) = P\left(0 < X < \frac{1}{2}\right) = \int_{x=0}^{x=1/2} f(x)dx$$
$$= \int_{x=0}^{x=1/2} 2x\,dx$$

Thinking of dx as a small change in X, and noting that $x = y/2$, then $dx = dy/2$. Substituting this into the integral above, we have

$$P(0 < Y < 1) = \int_{y/2=0}^{y/2=1/2} 2\left(\tfrac{1}{2}y\right)\left(\tfrac{1}{2}dy\right) = \int_{y=0}^{y=1} \tfrac{1}{2}y\,dy$$

The adjustment factor 1/2 that we obtained intuitively appears here in the relation of dx to dy. The mathematical name for this adjustment factor is the Jacobian of the transformation (actually its absolute value, as we will soon see). Its purpose is to make the integral expression in terms of x equal to that in terms of y. Now we are ready to describe the change-of-variable technique more precisely.

Let X be a continuous random variable with *pdf* $f(x)$. Let $Y = g(X)$ be a function that is strictly increasing or strictly decreasing. This condition ensures that the function is one-to-one, so that there is exactly one Y value for each X value and exactly one X value for each Y value. The importance of this condition on $g(X)$ is that we can solve $Y = g(X)$ for X. That is, we can find an inverse function $X = w(Y)$. Then the *pdf* for Y is given by

$$h(y) = f[w(y)] \cdot \left|\frac{dw(y)}{dy}\right| \tag{B.41}$$

where $|\,|$ denotes the absolute value.

Change of Variable Technique to Find the *pdf* of Y: Step by Step

1. Solve $y = g(x)$ for x in terms of y;
2. Substitute this for x in $f(x)$; and
3. Multiply by the absolute value of the derivative $dw(y)/dy$, which is called the Jacobian of the transformation.

The scale factor $|dw(y)/dy|$ is the adjustment factor that makes the probabilities (i.e., the integrals) come out right. In Example B.12 the inverse function is $X = w(Y) = Y/2$. The Jacobian term is $dw(y)/dy = d(y/2)/dy = \frac{1}{2}$, and $\left|dw(y)/dy\right| = \left|\frac{1}{2}\right| = \frac{1}{2}$.

EXAMPLE B.13 | Change of Variable: Continuous Case

Let X be a continuous random variable with *pdf* $f(x) = 2x$ for $0 < x < 1$. Let $Y = g(X) = 8X^3$ be the function of X in which we are interested. The function $Y = g(X) = 8X^3$ is strictly increasing for the set of values that X can take, $0 < x < 1$. The corresponding set of values that Y can take is $0 < y < 8$. Because the function is strictly increasing, we can solve for the inverse function

$$x = w(y) = \left(\frac{1}{8}y\right)^{1/3} = \frac{1}{2}y^{1/3}$$

and

$$\frac{dw(y)}{dy} = \frac{1}{6}y^{-2/3}$$

Applying the change-of-variable formula (B.41), we have

$$\begin{aligned} h(y) &= f\left[w(y)\right] \times \left|\frac{dw(y)}{dy}\right| \\ &= 2\left(\frac{1}{2}y^{1/3}\right) \times \left|\frac{1}{6}y^{-2/3}\right| \\ &= \frac{1}{6}y^{-1/3}, 0 < y < 8 \end{aligned}$$

The change-of-variable technique can be modified for the case of several random variables, X_1, X_2 being transformed into Y_1, Y_2. For a description of the method, which requires matrix algebra, see William Greene (2018) *Econometric Analysis*, 8th edition, Pearson Prentice Hall, pp. 1120–1121.

B.2.6 Truncated Random Variables

A truncated random variable is one whose probability density function is cutoff above or below some specified point. That is suppose that X is a continuous random variable such that $-\infty < x < \infty$ and its *pdf* is $f(x)$. The *pdf* $f(x)$ has the properties (i) $f(x) \geq 0$ and (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Now suppose that the underlying experiment is such that only x values greater than some value a are possible. What is the probability density function of this random variable? It is not simply $f(x)$ for $x > c$ because the *pdf* would not satisfy condition (ii) above, the area beneath it, which represents probability, would not total one. There is a simple fix-up. The density of a truncated random variable, such that $x > c$, is

$$f(x|x > c) = \frac{f(x)}{P(X > c)}$$

The adjustment makes the area equal to one.

Intuitively, what will happen to the expected value and variance of the truncated random variable, relative to the untruncated one? Thinking about it for a moment you can see that $E(X|x > c) > E(X)$ and $\text{var}(X|x > c) < \text{var}(X)$. Specific examples of truncated random variables will appear in the case of Poisson random variables (Section B.3.3) and normally distributed random variables (Section B.3.5).

B.3 Some Important Probability Distributions

In this section, we give brief descriptions and summarize the properties of the probability distributions used in this book.

B.3.1 The Bernoulli Distribution

Let the random variable X denote an experimental outcome with only two possible outcomes, A or B. Let $X = 1$ if the outcome is A and let $X = 0$ if the outcome is B. Let the probabilities of the outcomes be $P(X = 1) = p$ and $P(X = 0) = 1 - p$ where $0 \leq p \leq 1$. X is said to have a **Bernoulli distribution**. The *pdf* of this Bernoulli random variable is

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.42}$$

The expected value of X is $E(X) = p$, and its variance is $\text{var}(X) = p(1 - p)$. This random variable arises in choice models, such as the linear probability model (Chapters 7, 8, and 16) and in binary and multinomial choice models (Chapter 16).

B.3.2 The Binomial Distribution

If $X_1, X_2, \ldots, X_n$ are independent random variables, each having a Bernoulli distribution with parameter p, then $X = X_1 + X_2 + \cdots + X_n$ is a discrete random variable that is the number of successes (i.e., Bernoulli experiments with outcome $X_i = 1$) in n trials of the experiment. The random variable X is said to have a **binomial distribution**. The *pdf* of this random variable is

$$P(X = x|n, p) = f(x|n, p) = \binom{n}{x} p^x(1-p)^{n-x} \quad \text{for} \quad x = 0, 1, \ldots, n \tag{B.43}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the number of combinations of n things taken x at a time. This distribution has two parameters, n and p, where n is a positive integer indicating the number of experimental trials and $0 \leq p \leq 1$. These probabilities are tedious to compute by hand, but econometric software has functions to carry out the calculations. The discrete probabilities are illustrated in Figure B.3.

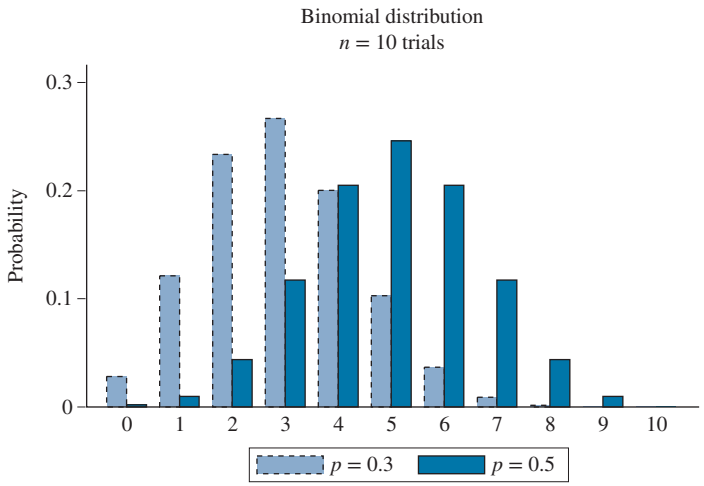


FIGURE B.3 **Binomial distributions for $n = 10$.**

The expected value and variance of X are

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p)$$

A related random variable is $Y = X/n$, which is the proportion of successes in n trials of an experiment. Its mean and variance are $E(Y) = p$ and $\text{var}(Y) = p(1-p)/n$.

B.3.3 The Poisson Distribution

Whereas a **binomial random variable** is the number of event occurrences in a given number of experimental trials, n, the Poisson random variable is the number of event occurrences in a given interval of time or space. The probability density function for this discrete random variable X is

$$P(X = x|\mu) = f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{for} \quad x = 0, 1, 2, 3, \ldots \tag{B.44}$$

Probabilities depend on the parameter μ, and $e \cong 2.71828$ is the base of natural logarithms. The expected value and variance of X are $E(X) = \text{var}(X) = \mu$. The **Poisson distribution** is used in models involving count variables (Chapter 16), such as the number of visits a person makes to a physician during a year. Probabilities for $x = 0$ to 10 for distributions with $\mu = 3$ and $\mu = 4$ are shown in Figure B.4.

In applications of count data, we sometimes only observe positive outcomes. For example, suppose we might survey individuals at a shopping mall and ask "How many times have you visited the mall this year?" The answer must be one or more. Using the notion of a truncated random variable introduced in Section B.2.6, the probability function in (B.44) becomes

$$f(x|\mu, x > 0) = \frac{f(x|\mu)}{P(X > 0)}$$

In the case of the Poisson distribution $P(X > 0) = 1 - P(X = 0) = 1 - e^{-\mu}$. Then the **truncated Poisson distribution** is

$$f(x|\mu, x > 0) = \frac{f(x|\mu)}{1 - P(X = 0)} = \frac{(e^{-\mu}\mu^x)/x!}{1 - e^{-\mu}} \quad \text{for} \quad x = 1, 2, 3, \ldots$$

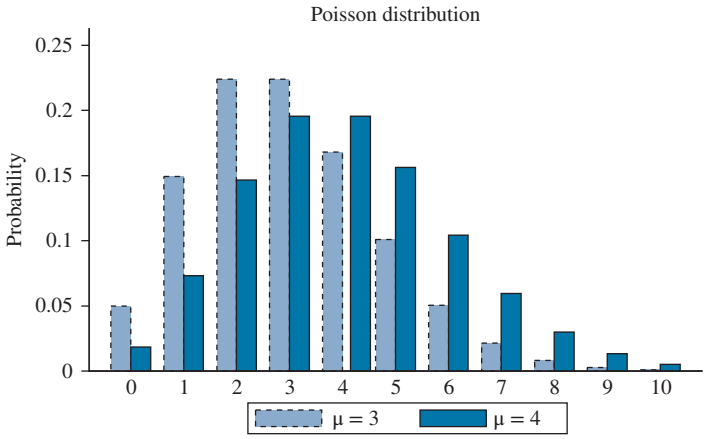


FIGURE B.4 **Poisson distributions.**

B.3.4 The Uniform Distribution

A continuous distribution that is vastly important for theoretical purposes is the **uniform distribution**. The random variable X with values $a \le x \le b$ has a uniform distribution if its *pdf* is given by

$$f(x|a,b) = \frac{1}{b-a} \quad \text{for} \quad a \le x \le b \tag{B.45}$$

The plot of the density function is given in Figure B.5.

The area under $f(x)$ between a and b is one, which is required of any probability density function for a continuous random variable. The expected value of X is the midpoint of the interval $[a, b]$, $E(X) = (a + b)/2$. This can be deduced from the symmetry of the distribution. The variance of X is $\text{var}(X) = E(X^2) - \mu^2 = (b - a)^2/12$.

An interesting special case occurs when $a = 0$ and $b = 1$, so that $f(x) = 1$ for $0 \le x \le 1$. The distribution, shown in Figure B.6, describes one common meaning of "a random number between zero and one."

The uniform distribution has the property that any two intervals of equal width have the same probability of occurring. That is,

$$P(0.1 \le X \le 0.6) = P(0.3 \le X \le 0.8) = P(0.21131 \le X \le 0.71131) = 0.5$$

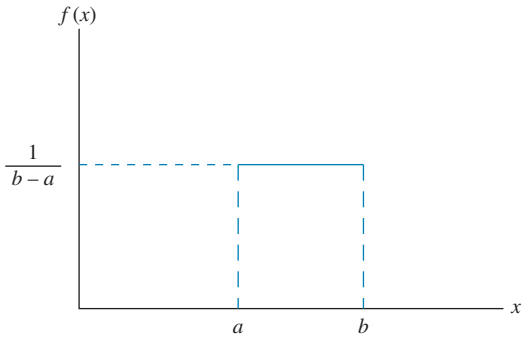


FIGURE B.5 **A uniform distribution.**

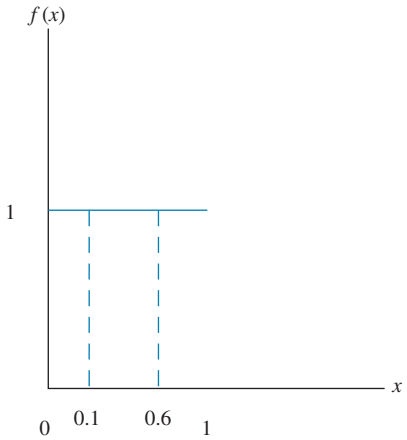


FIGURE B.6 **A uniform distribution on [0, 1] interval.**

Picking a number randomly between zero and one is conceptually complicated by the fact that the interval has an uncountably infinite number of values, and the probability of any one of them occurring is zero. What is more likely meant by such a statement is that each interval of equal width has the same probability of occurring, no matter how narrow. This is exactly the nature of the uniform distribution.

B.3.5 The Normal Distribution

The normal distribution was described in the Probability Primer, Section P.6. A point not stressed at that time was why we must consult tables, like Statistical Table 1 to calculate normal probabilities. For example, we now know that for the continuous and normally distributed random variable X, with mean μ and variance σ^2, the probability that X falls in the interval $[a, b]$ is

$$\int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-\mu)^2/2\sigma^2\right] dx$$

Unfortunately this integral does not have a closed-form algebraic solution. Consequently, we wind up working with tabled values containing numerical approximations to areas under the **standard normal distribution**, or we use computer software functions in a similar manner.

Moments of the Normal Distribution

If X is a random variable, then $E(X^r)$ is called the rth moment of the random variable about the origin. Sometimes they are called *raw* moments. If $X \sim N(\mu,\sigma^2)$, then we have the following useful expressions for the first three moments about the origin:

$$E(X) = \mu$$
$$E(X^2) = \mu^2 + \sigma^2$$
$$E(X^3) = 0$$

For any random variable X, $E(X-\mu)^r$ is the rth moment of the random variable about its mean. Sometimes, these are called *central* moments. For the normal random variable $X \sim N(\mu,\sigma^2)$, these are

$$E(X-\mu) = 0$$
$$E\left[(X-\mu)^2\right] = \sigma^2$$
$$E\left[(X-\mu)^3\right] = 0$$
$$E\left[(X-\mu)^4\right] = 3\sigma^4$$

The second moment about the mean $E\left[(X-\mu)^2\right] = \sigma^2$ is the variance of the random variable. The third moment, $E\left[(X-\mu)^3\right] = 0$, is related to the *skewness* of the probability density function. Because the normal distribution is symmetrical, it is not skewed, its skewness is zero. It is also true that all odd central moments are zero, so that $E\left[(X-\mu)^r\right] = 0$ if r is an odd number. The fourth moment about the mean, $E\left[(X-\mu)^4\right] = 3\sigma^4$, is related to the *kurtosis* of the distribution, which is a measure of the thickness of the tails of the distribution. For the normal distribution, the standardized fourth moment $E\left[(X-\mu)^4/\sigma^4\right] = 3$ is a useful reference point for tail thickness. For more about population moments see Appendix C.4.

The Truncated Normal Distribution

In Section B.2.6, we introduced the notion of a truncated random variable. The truncated normal distribution has been studied quite intensely. Suppose that $X \sim N(\mu,\sigma^2)$ but the distribution is **truncated from below** so that $x > c$. Then

$$f(x|x > c) = \frac{f(x)}{P(X > c)}$$

For the normal distribution

$$P(X > c) = P\left(\frac{X-\mu}{\sigma} > \frac{c-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{c-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

where $\Phi(\alpha)$ is the cumulative distribution function of the standard normal random variable evaluated at $\alpha = (c-\mu)/\sigma$. Then

$$f(x|x > c) = \frac{f(x)}{1 - \Phi(\alpha)}$$

Following Greene (2018, p. 921), define the **Inverse Mill's Ratio** as

$$\lambda(\alpha) = \begin{cases} \dfrac{\phi(\alpha)}{1-\Phi(\alpha)} & \text{if truncation is from below, so that } x > c \\ \dfrac{-\phi(\alpha)}{\Phi(\alpha)} & \text{if truncation is from above, so that } x < c \end{cases}$$

where $\phi(\alpha)$ is the probability density function of the standard normal random variable evaluated at $\alpha = (c-\mu)/\sigma$. Then the expected value of the truncated normal random variable is

$$E(X|\text{truncation}) = \mu + \sigma\lambda(\alpha)$$

Letting $\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$, the variance of the truncated normal random variable is

$$\text{var}(X|\text{truncation}) = \sigma^2[1 - \delta(\alpha)]$$

This is consistent with the intuition about the variance of a truncated variable in Section B.2.6 because $0 < \delta(\alpha) < 1$.

The normal distribution is related to the chi-square, *t*-, and *F*-distributions, which we now discuss.

B.3.6 The Chi-Square Distribution

Chi-square random variables arise when standard normal random variables are squared. If Z_1, Z_2, …, Z_m denote m *independent* $N(0,1)$ random variables, then

$$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2 \sim \chi^2_{(m)} \tag{B.46}$$

The notation $V \sim \chi^2_{(m)}$ is read as: The random variable V has a chi-square distribution with m **degrees of freedom**. The degrees of freedom parameter m indicates the number of *independent* $N(0,1)$ random variables that are squared and summed to form V. The value of m determines the entire shape of the **chi-square distribution**, including its mean and variance as

$$E(V) = E\left[\chi^2_{(m)}\right] = m$$

$$\text{var}(V) = \text{var}\left[\chi^2_{(m)}\right] = 2m \tag{B.47}$$

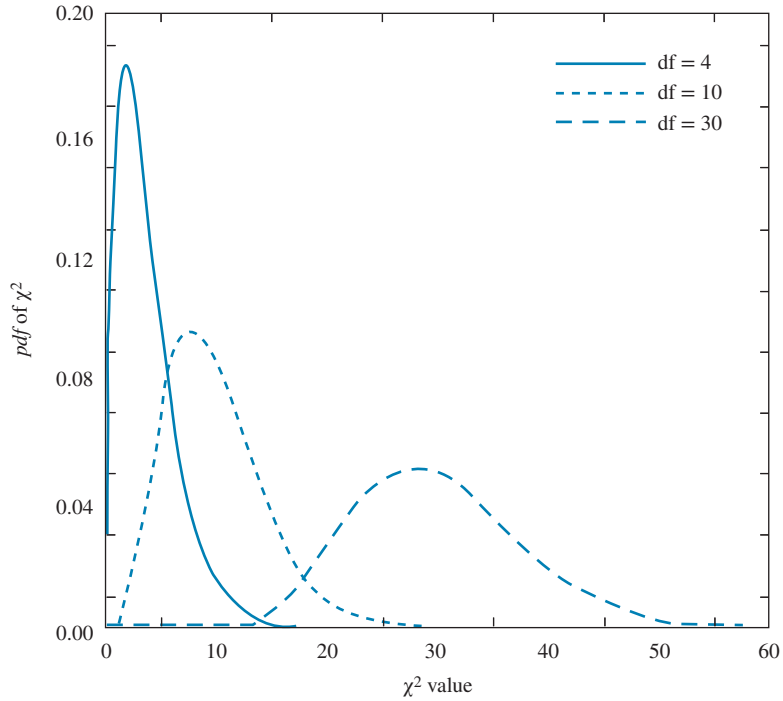


FIGURE B.7(a) **The chi-square distribution.**

In Figure B.7(a) graphs of the chi-square distribution for various degrees of freedom are presented. The values of V must be nonnegative, $v \geq 0$, because V is formed by squaring and summing m standardized normal, $N(0,1)$, random variables. The distribution has a long tail, or is *skewed*, to the right. As the degrees of freedom m gets larger, however, the distribution becomes more symmetric and "bell-shaped." In fact, as m gets larger, the chi-square distribution converges to, and essentially becomes, a normal distribution.

The 90th, 95th, and 99th percentile values of the chi-square distribution for selected values of the degrees of freedom are given in Statistical Table 3. These values are often of interest in hypothesis testing.

In the definition (B.46) of the chi-square random variable the Z_i, $i = 1, \ldots, m$ are statistically independent standard normal, $N(0, 1)$, random variables. If, instead, V is equal to the sum of squares of normal random variables $(Z_i + \delta_i)$ that have a non-zero mean δ_i and variance 1, then V has a **non-central chi-square distribution** with m degrees of freedom and **non-centrality parameter** $\delta = \delta_1^2 + \delta_2^2 + \cdots + \delta_m^2$, which is denoted by $\chi^2_{(m,\delta)}$. If all $\delta_i = 0$ then we have the usual **central chi-square** distribution. That is,

$$V = (Z_1 + \delta_1)^2 + (Z_2 + \delta_2)^2 + \cdots + (Z_m + \delta_m)^2 \sim \chi^2_{(m,\delta)}$$

In Figure B.7(b) we plot a few non-central chi-square distributions, all having $m = 10$ degrees of freedom.

The effect of the non-centrality parameter is to shift the chi-square density function to the right, increasing both the mean and the variance, which become $E\left[\chi^2_{(m,\delta)}\right] = m + \delta$ and $\text{var}\left[\chi^2_{(m,\delta)}\right] = 2(m + 2\delta)$.

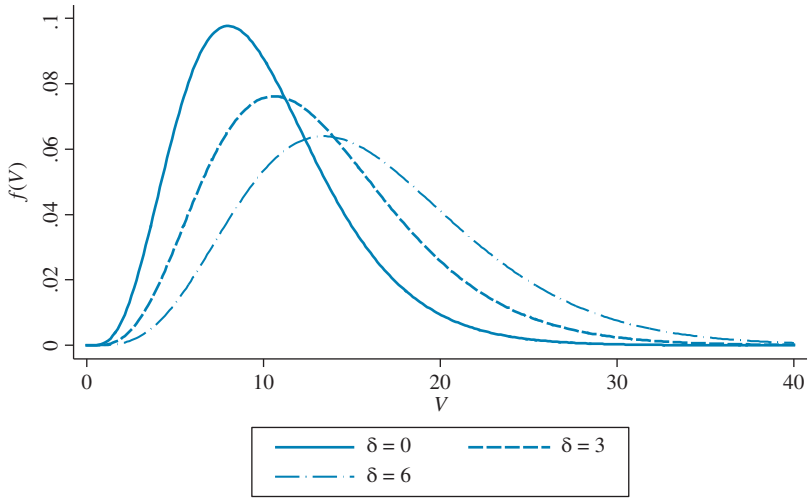


FIGURE B.7(b) **Non-central chi-square distributions, $m = 10$ degrees of freedom and non-centrality $\delta = 0, 3, 6$.**

B.3.7 The *t*-Distribution

A t random variable (no upper case) is formed by dividing a standard normal random variable $Z \sim N(0,1)$ by the square root of an *independent* chi-square random variable, $V \sim \chi^2_{(m)}$, that has been divided by its degrees of freedom m. If $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$, and if Z and V are independent, then

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)} \tag{B.48}$$

The t-distribution's shape is completely determined by the degrees of freedom parameter, m, and the distribution is symbolized by $t_{(m)}$.

Figure B.8(a) shows a graph of the t-distribution with $m = 3$ degrees of freedom relative to the $N(0,1)$. Note that the t-distribution is less "peaked," and more spread out than the $N(0,1)$. The t-distribution is symmetric, with mean $E(t_{(m)}) = 0$ and variance $\text{var}(t_{(m)}) = m/(m-2)$. As the degrees of freedom parameter $m \to \infty$, the $t_{(m)}$ distribution approaches the standard normal $N(0,1)$.

Computer programs have functions for the *cdf* of t-random variables that can be used to calculate probabilities. Since certain probabilities are widely used, Statistical Table 2 contains frequently used percentiles of t-distributions, called **critical values** of the distribution. For example, the 95th percentile of a t-distribution with 20 degrees of freedom is $t_{(0.95,20)} = 1.725$. The t-distribution is symmetric, so Statistical Table 2 shows only the right tail of the distribution.

The statistic formed from a $N(\delta,1)$ random variable and an independent central chi-square random variable with m degrees of freedom is called a **non-central *t*-random variable**,

$$t = \frac{Z + \delta}{\sqrt{V/m}} \sim t_{(m,\delta)}$$

This distribution has two parameters, the degrees of freedom, m, and the **non-centrality parameter** δ. The usual t-random variable in (B.48) has non-centrality parameter $\delta = 0$ and is sometimes called the **central *t*-distribution**. The additive factor in the numerator causes the resulting distribution to be centered at a value other than zero if $\delta \neq 0$. In Figure B.8(b), we plot the $t_{(3,\delta)}$

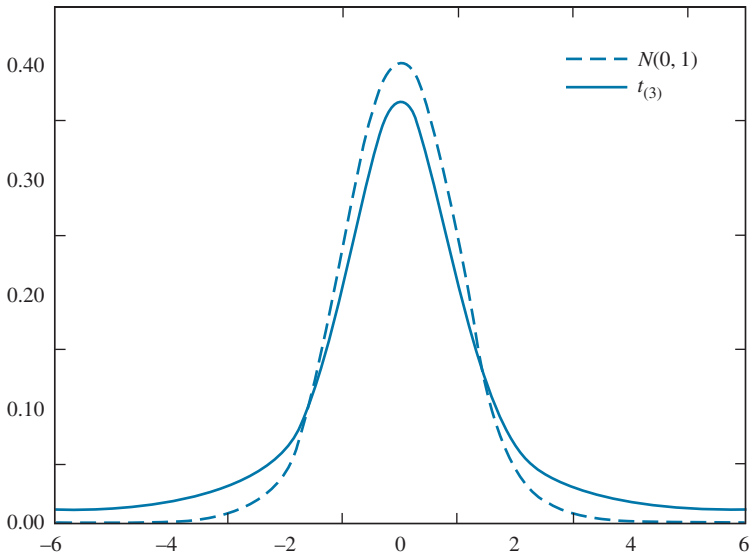


FIGURE B.8(a) **The standard normal and $t_{(3)}$ probability density functions.**

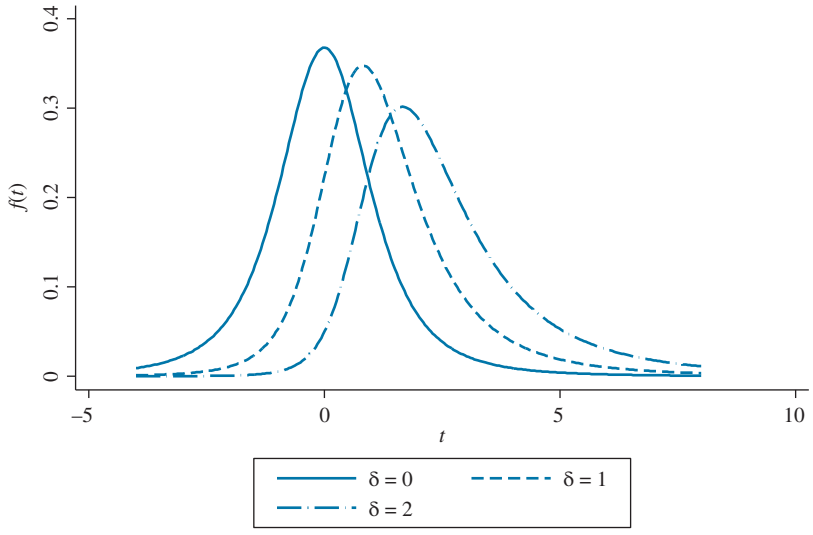


FIGURE B.8(b) **Non-central t-distributions, $m = 3$ degrees of freedom and non-centrality $\delta = 0, 1, 2$.**

density for values of $\delta = 0, 1, 2$. The positive non-centrality parameter shifts the density function rightward.

B.3.8 The F-Distribution

An F-random variable is formed by the ratio of two independent chi-square random variables that have been divided by their degrees of freedom. If $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$, and if V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)} \tag{B.49}$$

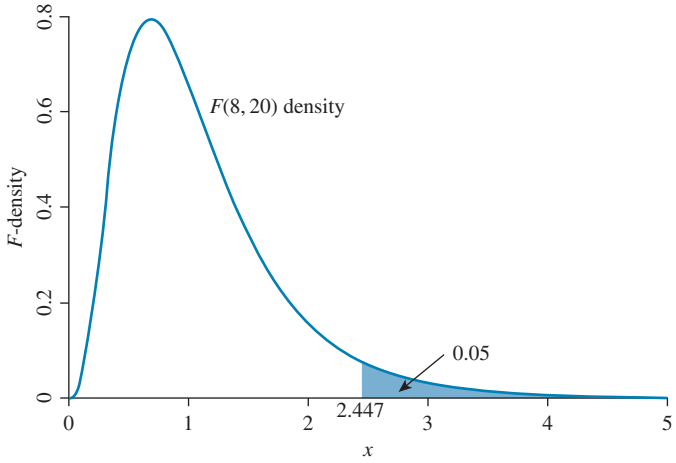


FIGURE B.9(a) **The 95th percentile of an $F_{(8,20)}$-random variable.**

The F-distribution is said to have m_1 ***numerator degrees of freedom*** and m_2 ***denominator degrees of freedom***. The values of m_1 and m_2 determine the shape of the distribution, which in general looks like Figure B.9(a). The range of the random variable is $(0,\infty)$, and it has a long tail to the right. For example, the 95th percentile value for an F-distribution with $m_1 = 8$ numerator degrees of freedom and $m_2 = 20$ denominator degrees of freedom is $F_{(0.95,8,20)} = 2.447$. Critical values (two decimal places) for the ***F*-distribution** are given in Statistical Table 4 (the 95th percentile) and Statistical Table 5 (the 99th percentile).

In the definition (B.49), the numerator chi-square random variable V_1 has a **central chi-square** distribution, with non-centrality parameter $\delta = 0$. The central and non-central chi-square distributions are discussed in Section B.3.6. If the numerator in (B.49) has a non-central chi-square distribution, $V_1 \sim \chi^2_{(m_1,\delta)}$ with m_1 degrees of freedom and non-centrality, δ, then the F-random variable has a **non-central** F-distribution with numerator degrees of freedom m_1, denominator degrees of freedom m_2 and non-centrality parameter δ. This distribution is denoted by $F_{(m_1,m_2,\delta)}$. In Figure B.9(b), we show several density functions for comparison with

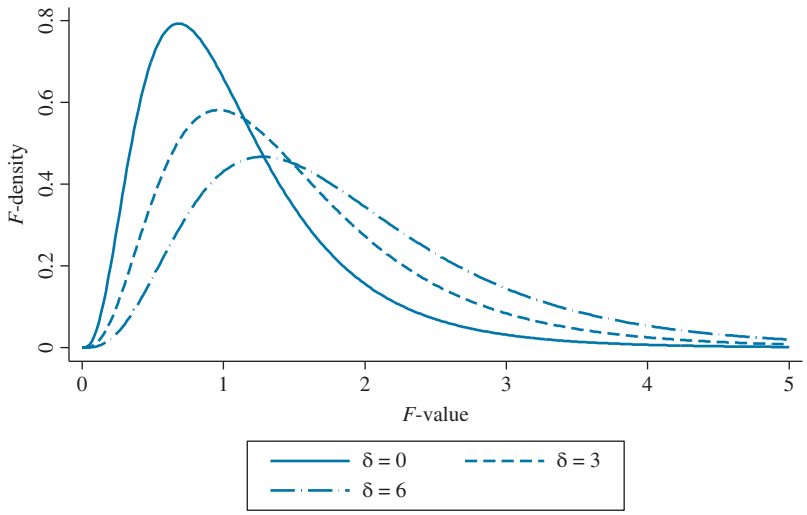


FIGURE B.9(b) **Non-central $F_{(8,\,20,\,\delta)}$-distributions with $\delta = 0,\ 3,\ 6$.**

Figure B.9(a). These have degrees of freedom $m_1 = 8$, $m_2 = 20$, and non-centrality $\delta = 0,\ 3,\ 6$. As the non-centrality parameter increases, the F-density moves to the right, increasing both its mean and variance.

B.3.9 The Log-Normal Distribution

A continuous random variable X is said to have a **log-normal** distribution if

$$\ln(X) \sim N(\mu, \sigma^2), \quad x > 0$$

The probability density function of X is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{[\ln(x) - \mu]^2}{2\sigma^2}\right\}, \quad x > 0$$

Probabilities are computed using the *cdf* of the standard normal random variable, $\Phi(z)$. That is

$$P(X \le c) = P[\ln(X) \le \ln(c)] = P\left\{[(\ln(X) - \mu)/\sigma] \le [(\ln(c) - \mu)/\sigma]\right\}$$

$$= \Phi[(\ln(c) - \mu)/\sigma]$$

The parameters μ and σ^2 are the mean and variance of $\ln(X)$. The *pdf* of X is not symmetrical. The **median** of X is $m = \exp(\mu)$ and $\mu = \ln(m)$.[1] The expected value of X is

$$E(X) = m\exp(\sigma^2/2) = \exp(\mu)\exp(\sigma^2/2) = \exp(\mu + \sigma^2/2)$$

Using $\omega = \exp(\sigma^2)$, the variance of X is

$$\text{var}(X) = m^2\omega(\omega - 1) = \exp(2\mu)\exp(\sigma^2)[\exp(\sigma^2) - 1] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$$

The mode of the density is m/ω so that $E(X) = \text{mean} > \text{median} > \text{mode}$. In Figure B.10, we plot the log-normal density for several choices of σ with median $m = 1$.

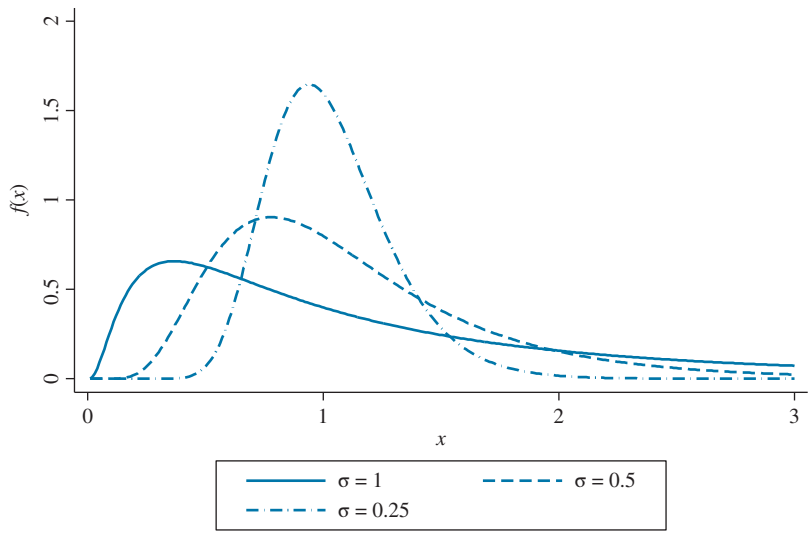


FIGURE B.10 **Log-normal densities. With median $m = 1$ and shape $\sigma = 1, 0.5, 0.25$.**

[1] In the statistics literature σ is sometimes called the **shape** parameter and m the **scale** parameter.

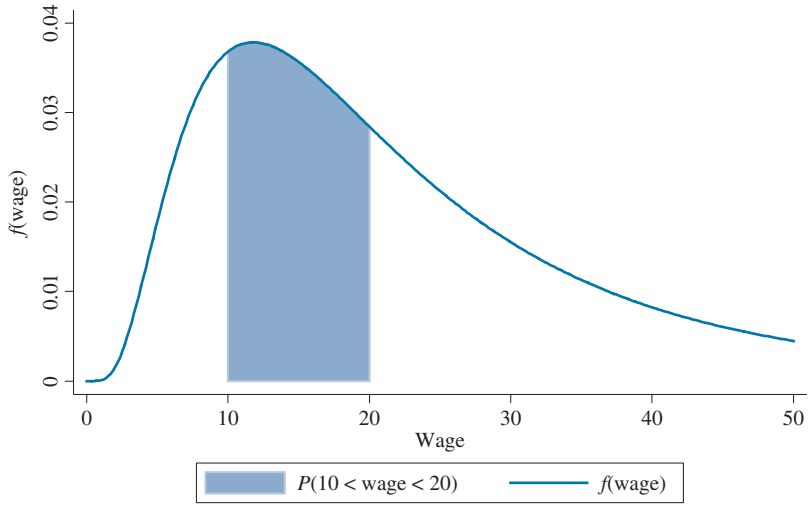


FIGURE B.11 **Hypothetical probability density function for *WAGE*, log-normal with $m = 19.23$ and $\sigma = 0.7$.**

A common use of the log-normal distribution in Economics is for wages, incomes, and house prices. These variables are positive, and the distributions are skewed with a long tail to the right, indicating that a small portion of the population has large values. Using the data file *cps5*, the median wage is \$19.23, and the mean wage is \$23.5. Using the expression for the expected value of a log-normal distribution $E(X) = m\exp\left(\sigma^2/2\right)$, we can calculate the shape parameter $\sigma = \sqrt{2\ln(E(X)/m)}$, which is about 0.7 using the *cps5* data values. Then the implied distribution of *WAGE* is shown in Figure B.11. What is the probability that a randomly chosen worker will have an hourly wage between \$10 and \$20? Graphically, it is the area under the *pdf* between 10 and 20. The calculated probability, using our approximated log-normal distribution is

$$P(10 < WAGE < 20) = \Phi\left[\frac{\ln(20) - \ln(19.23)}{0.7}\right] - \Phi\left[\frac{\ln(10) - \ln(19.23)}{0.7}\right]$$

$$= \Phi(0.05609) - \Phi(-0.93412)$$

$$= 0.52236 - 0.17512 = 0.34724$$

In the *cps5* data, 38.95% of the individuals have a wage between \$10 and \$20, so our rough approximation using the log-normal distribution is not far off.

B.4 Random Numbers

In several chapters we carry out Monte Carlo simulations to illustrate the sampling properties of estimators. See, for example, Chapters 3, 4, 5, 10, 11, and 16. To use Monte Carlo simulations we rely upon the ability to create **random numbers** from specific probability distributions, such

as the uniform and the normal. Using computer simulations is widespread in all sciences. In this section we introduce to you this aspect of computing.[2] You should first realize that the idea of creating random numbers using a computer is paradoxical, because by definition random numbers that are "created" cannot be truly random. The random numbers generated by a computer are **pseudo-random numbers** in that they "behave as if they were random." We present one method for generating pseudo-random numbers called the **inverse transformation** approach, or the **inversion method**. This method assumes that we have the ability to generate pseudo-random numbers from the **uniform distribution** (see Sections B.3.4 and B.4.1) on the (0, 1) interval. The uniformly distributed random variables are then transformed into random variables with other distributions.

EXAMPLE B.14 | An Inverse Transformation

Let U be a random variable with a uniform distribution. It is a continuous random variable with *pdf* $h(u) = 1$ for $u \in (0, 1)$. See Figure B.6 for an illustration. If $Y = U^{1/2}$, then $0 < y < 1$. Furthermore, the square root function is strictly increasing, so that we can apply the change-of-variable technique to find the *pdf* of Y. The inverse function is $U = w(Y) = Y^2$, and the Jacobian of the transformation is $dw(y)/dy = d\left(y^2\right)/dy = 2y$. The *pdf* of Y is then

$$f(y) = h\left[w(y)\right] \times \frac{dw(y)}{dy} = 1 \times 2y = 2y, \quad 0 < y < 1 \tag{B.50}$$

This is a distribution that we have used in Examples B.12 and B.13. The importance of this example is that it shows that we can obtain a random number from the distribution in (B.50) by taking the square root of a random number from a uniform distribution.

Example B.14 leads us toward a general technique, the inversion method, for drawing random numbers from certain distributions. Suppose you wish to obtain a random number from a specific probability distribution, with *pdf* $f(y)$ and *cdf* $F(y)$.

The Inversion Method: Step by Step

1. Obtain a uniform random number u_1 in the (0, 1) interval.
2. Let $u_1 = F\left(y_1\right)$.
3. Solve the equation in step 2 for $y_1 = F^{-1}\left(u_1\right)$.
4. The value y_1 is a random number from the *pdf* $f(y)$.

The inversion method can be used to draw random numbers from any distribution that permits you to carry out step 3. The solution is often denoted $y_1 = F^{-1}\left(u_1\right)$, where F^{-1} is called the **inverse cumulative distribution function**. The *cdf* function F is said to be **invertible**.

[2] A well-written book on the subject is by James E. Gentle (2003) *Random Number Generation and Monte Carlo Methods,* New York: Springer. Also, J. F. Kiviet (2011) Monte Carlo Simulation for Econometricians, *Foundations and Trends® in Econometrics*, vol 5, nos 1–2.

EXAMPLE B.15 | The Inversion Method: An Example

Suppose the target distribution, from which we want a random number, is $f(y) = 2y, 0 < y < 1$. The *cdf* of Y is $P(Y \leq y) = F(y) = y^2, 0 < y < 1$. The two distributions are shown in Figure B.12. Set a uniform random number $u_1 = F(y_1) = y_1^2$ and solve to obtain $y_1 = F^{-1}(u_1) = (u_1)^{1/2}$. The value y_1 is a random value, or a **random draw**, from the probability distribution $f(y) = 2y, 0 < y < 1$. This agrees perfectly with the result in Example B.6, where we showed that the square root of a uniform random variable has this *pdf*.

In Figure B.12(a), suppose the uniform random number value is $u_1 = 0.16$. It falls between 0 and 1, along the vertical axis of the *cdf* function $F(y)$. The value $u_1 = 0.16$ corresponds to the value $y_1 = 0.4 = (u_1)^{1/2} = (0.16)^{1/2}$ on the horizontal axis. In the lower panel we see the connection between the *pdf* and the *cdf*. The area under the *pdf* to the left of $y_1 = 0.4$ is the probability $P(0 < Y < 0.4) = 0.16$. For every randomly drawn uniform random number u_i there is a unique corresponding y_i from the distribution $f(y) = 2y, 0 < y < 1$.

To illustrate, in the data file *uniform1*, we have 1,000 observations on two independent uniform random variables *U1* and *U2*.[3] Figure B.13 shows the histogram of *U1*. There are 10 intervals and approximately 10% of the values fall into each, as we would expect for values from a uniform distribution.

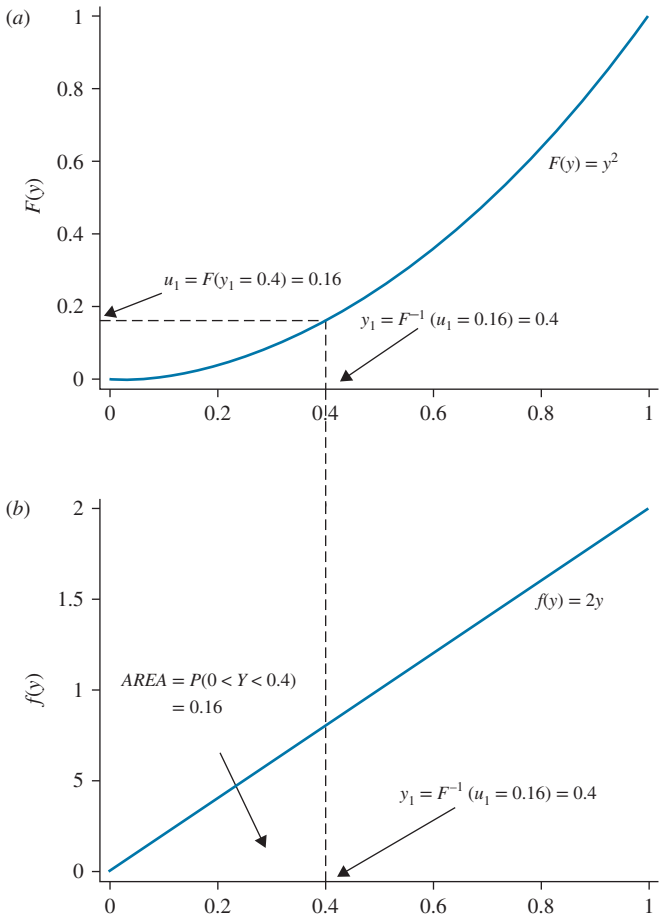


FIGURE B.12 **(*a*) Cumulative distribution function and (*b*) probability density function.**

[3]The data file *uniform2* contains 10,000 observations if you prefer a larger sample.

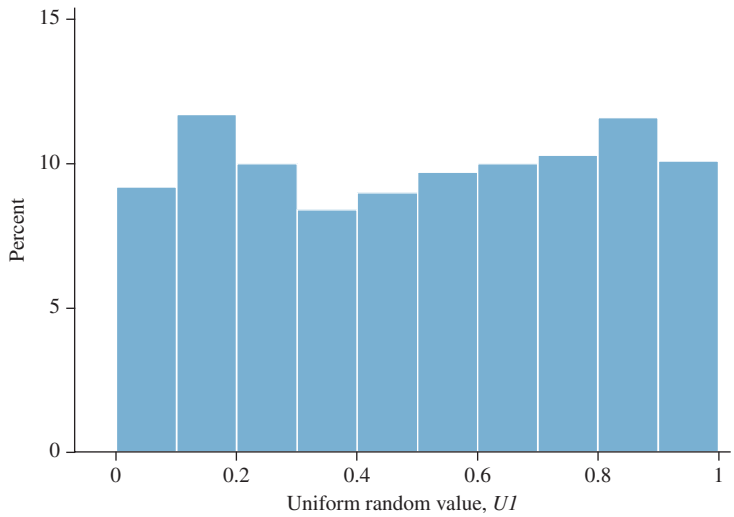


FIGURE B.13 **Histogram of 1,000 uniform random values.**

Let *Y1* be the square root of the *U1* values. The histogram of these values is shown in Figure B.14. It looks like a triangle, doesn't it? Just like the density $f(y) = 2y, 0 < y < 1$.

As a second example, let us consider a slightly more exotic distribution. The **extreme value distribution** is the foundation of logit choice models that are discussed in Chapter 16. It has probability density function $f(v) = \exp(-v)\exp(-\exp(-v))$, depicted in Figure B.15. The extreme value *cdf* is $F(v) = \exp(-\exp(-v))$. Despite its complicated-looking form, we can obtain values from this distribution using $v = F^{-1}(u) = -\ln(-\ln(u))$. Using the 1,000 values *U1* in data file *uniform1*, we obtain the histogram of values from the extreme value distribution shown in Figure B.16.[4] The solid curve superimposed on the histogram looks much like the extreme value density function in Figure B.15.

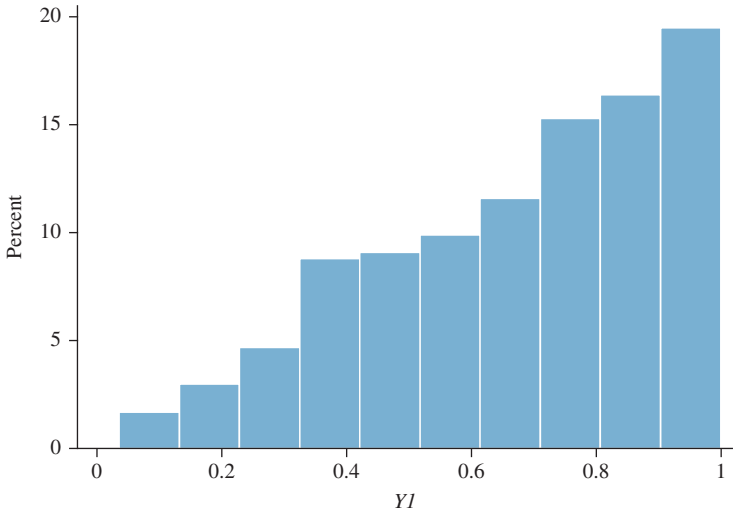


FIGURE B.14 **Histogram of 1,000 square roots of uniform random values.**

[4]The solid curve is a kernel density fitted to the data using a Gaussian kernel. See Appendix C.10 for more on kernel densities.

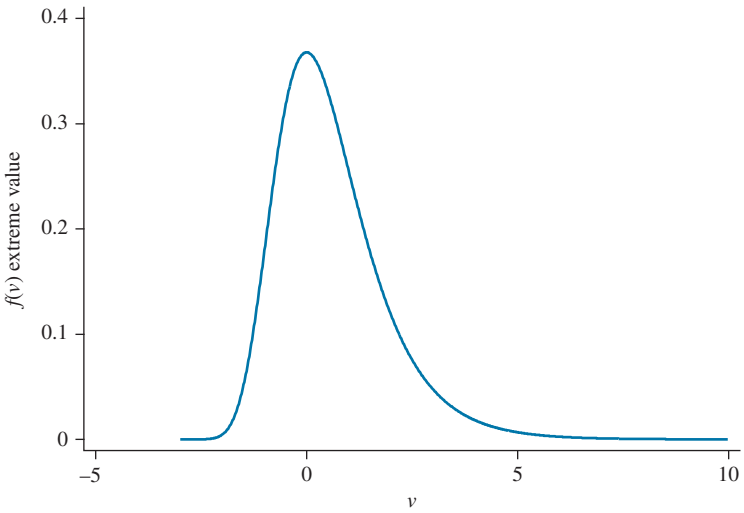


FIGURE B.15 **The extreme value distribution.**

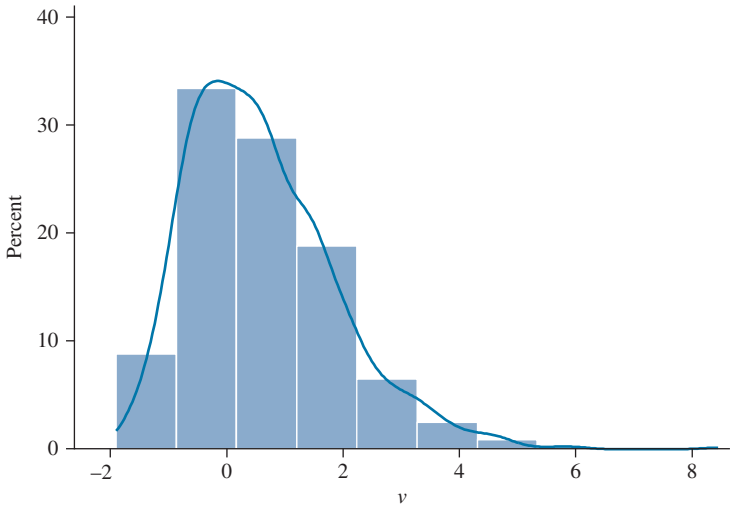


FIGURE B.16 **Histogram of simulated draws from the extreme value distribution.**

To summarize, the **inversion method** for generating random numbers from specific distributions depends upon (1) the ability to obtain uniform random numbers and (2) the distribution having a *cdf* that is invertible. The procedure does not work for joint distributions.

Knowing the inversion method, you can generate random variables from other distributions given a uniform random number generator. Books on statistical distributions[5] have instructions on how to transform uniform random numbers into a wide variety of distributions. A particular method for generating normal random numbers is illustrated in Exercise B.8.

[5]See, for example, Catherine Forbes, Merran Evans, Nicholas Hastings, Brian Peacock (2010) *Statistical Distributions*, 4th ed., John Wiley and Sons, Inc.

B.4.1 Uniform Random Numbers

The inversion method depends upon the ability to obtain random numbers from a uniform distribution. The generation of "random numbers" when used without modifiers usually means uniform random numbers, which is a field of study in and of itself. As noted earlier, the notion of computer-generated random numbers is illogical. Computers use algorithms to do their work; an algorithm is a formula so that the product is not "random," but randomlike. Computers generate **pseudo-random numbers**. Enter that term into your favorite search engine and you will find many, many links.

One bit of notation that appears in citations is for the mathematical **modulus**, denoted "a mod b." This is shorthand for the remainder resulting from dividing a by b. One method for calculating the modulus is[6]

$$n \bmod m = n - m\,\text{ceil}(n/m) + m \tag{B.51}$$

where **ceil** is short for the **ceiling** function that rounds up[7] to the next integer. To see how this works:

$$7 \bmod 3 = 1 = 7 - 3\,\text{ceil}(7/3) + 3 = 7 - 3\,\text{ceil}(2.3333) + 3 = 7 - 3 \cdot 3 + 3 = 1$$

A standard method for creating a uniform random number is the **linear congruential generator**.[8] Consider the recursive relationship

$$X_n = \left(aX_{n-1} + c\right) \bmod m \tag{B.52}$$

where a, c, and m are constants that we choose. It means that X_n takes the value equal to the remainder obtained by dividing $aX_{n-1} + c$ by m. It is a recursive relationship because the nth value depends on the $(n-1)$th. That means we must choose a starting value X_0, which is called the **random number seed**. Everyone using the same seed, and values a, c, and m will generate the same string of numbers. The value m is the divisor in (B.52), and it determines the maximum period of the recursively generated values. The uniform random values falling in the interval (0, 1) are obtained as $U_n = X_n/m$. The value of m is often chosen to be 2^{32} when using computers with 32-bit architecture. The values of a and c are critical to the success of the random number generator. Bad choices result in sequences of numbers that are not random. For example, type RANDU into your search engine. This was a popular random number generator in the 1960s (I used it too!) that was later discovered to be very flawed, failing tests of randomness.[9]

EXAMPLE B.16 | Linear Congruential Generator Example

To illustrate that the process defined in (B.52) can generate apparently random numbers, we choose $X_0 = 1234567$, $a = 1664525$, $b = 1013904223$, and $m = 2^{32}$ and create 10,000 data values, labeled *U1* in the data file *uniform3*.[10] Using a histogram with 20 bins, we would expect 5% of the values in each, and as Figure B.17 illustrates, that is about what we get.

The 10,000 values for *U1* have sample mean 0.4987197 and variance 0.0820758 compared to the true mean and variance for a uniform distribution of 0.5 and 0.08333. The minimum and maximum values are 0.0000327 and 0.9998433, respectively.

[6] www.functions.wolfram.com/IntegerFunctions/Mod/27/01/03/01/0001/.

[7] ceil(x) is the smallest integer not less than x.

[8] A description and link to sources is www.en.wikipedia.org/wiki/Linear_congruential_generator.

[9] George Marsaglia developed a series of tests for randomness that are widely used. They are available at www.stat.fsu.edu/pub/diehard/.

[10] The variable *U2* in this file uses seed 987654321.

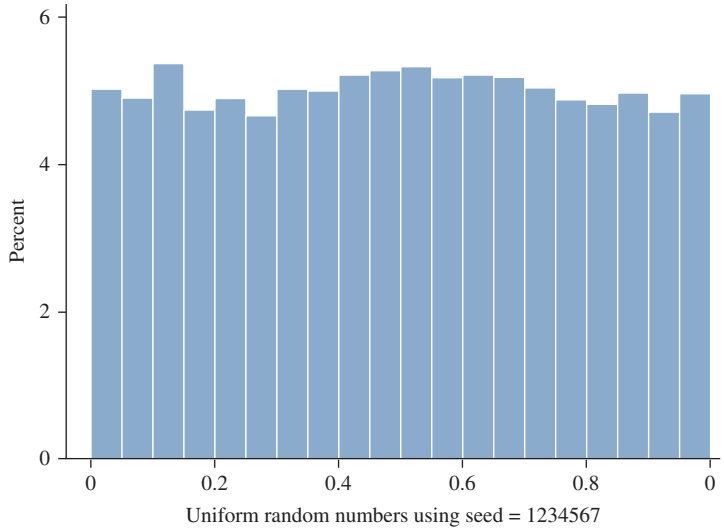


FIGURE B.17 **Histogram of 10,000 generated uniform random values.**

The lessons learned from these exercises are that random numbers are not random, and some random number generators are better than others. Ones that are popularly cited are the Mersenne twister and the KISS+Monster algorithm. New ones continue to be developed, and each software provider uses different algorithms which are predominately kept secret, or difficult to discover at any rate.

The third lesson is that you should probably **not** attempt to write your own random number algorithms. Professor Ken Train, an econometrician who has studied computational methods a great deal, says,[11] "From a practical perspective, my advice is the following: unless one is willing to spend considerable time investigating and resolving (literally, re-solving)…" the issues related to designing pseudo-random number routines "… it is probably better to use available routines rather than write a new one." Our advice is to use your software to generate random numbers, but when documenting your work, cite the software used and the software version, as revisions can change results from one version to another.

B.5 Exercises

B.1 Let $X_1, X_2, \ldots, X_n$ be independent random variables which all have the same probability distribution, with mean μ and variance σ^2. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a. Use the properties of expected values to show that $E\left(\overline{X}\right) = \mu$.

b. Use the properties of variance to show that $\text{var}\left(\overline{X}\right) = \sigma^2/n$. How have you used the assumption of independence?

[11] *Discrete Choice Methods with Simulation*, 2nd ed., 2009, Cambridge University Press, p. 206.

B.2 Suppose that Y_1, Y_2, Y_3 is a sample of observations from a $N(\mu, \sigma^2)$ population but that Y_1, Y_2, and Y_3 are *not* independent. In fact, suppose that

$$\text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_3) = \text{cov}(Y_1, Y_3) = \frac{\sigma^2}{2}$$

Let $\overline{Y} = (Y_1 + Y_2 + Y_3)/3$.

a. Find $E(\overline{Y})$.

b. Find $\text{var}(\overline{Y})$.

B.3 Let X be a continuous random variable with probability density function given by

$$f(x) = -\frac{1}{2}x + 1, \quad 0 \le x \le 2$$

a. Graph the density function $f(x)$.

b. Find the total area beneath $f(x)$ for $0 \le x \le 2$.

c. Find $P(X \ge 1)$ using both geometry and integration.

d. Find $P(X \le \frac{1}{2})$.

e. Find $P(X = 1\frac{1}{2})$.

f. Find the expected value and variance of X.

g. Find the cumulative distribution function of X.

B.4 Let X be a uniform random variable on the interval (a, b).

a. Use integration techniques to find the mean and variance of X.

b. Find the cumulative distribution function of X.

B.5 Use the recursive relationship in (B.52) with $X_0 = 79$, $m = 100$, $a = 263$, and $c = 71$ to generate 40 values $X_1, X_2, \ldots, X_{40}$. Do the resulting numbers appear random? Is this a good random number generator, or not?

B.6 Let X have a normal distribution with mean μ and variance σ^2. Use the change-of-variable technique to find the probability density function of $Y = aX + b$.

B.7 Show that if $E(Y|X) = E(Y)$, then $\text{cov}[Y, g(X)] = 0$ for any function $g(X)$.

B.8 Normal random numbers are useful for Monte Carlo simulations. One way to generate them is using the Box–Muller transformation. The Box–Muller transformation creates two new random variables, *Z1* and *Z2*, that have independent $N(0, 1)$ distributions, using

$$Z1 = \sqrt{-2\ln(U1)}\cos(2\pi U2), \quad Z2 = \sqrt{-2\ln(U1)}\sin(2\pi U2)$$

a. Construct a histogram of *Z1* and *Z2* obtained by using the 1,000 uniform random values *U1* and *U2* in data file *uniform1* (or the 10,000 values in the data file *uniform2*). Is the distribution of values "bell shaped"?

b. Calculate the summary statistics for *Z1* and *Z2*. Are the sample mean and variance close to zero and one, respectively?

c. Construct a scatter diagram for *Z1* and *Z2*. That is, plot *Z1* (vertical axis) and *Z2* (horizontal axis) in the *x*–*y* plane. Is there any evidence of positive or negative correlation?

B.9 Let X be a continuous random variable with *pdf* $f(x) = 3x^2/8$ for $0 < x < 2$. Compute

a. $P(0 < X < \frac{1}{2})$

b. $P(1 < X < 2)$

B.10 A continuous random variable X is said to have an exponential distribution if its *pdf* is $f(x) = e^{-x}$, $x \ge 0$.

a. Plot this density function for $0 \le x \le 10$.

b. The cumulative distribution function for X is $F(x) = 1 - e^{-x}$. Plot this function over the interval $0 \le x \le 10$. Is it strictly increasing or decreasing, or are you unsure?

c. Use the inverse transformation method to draw random values *X1* from this distribution. Use the 1,000 values for *U1* in data file *uniform1* or the 10,000 values for *U1* in data file *uniform2*. Construct a histogram of the values you have created. Does it resemble the plot in (a)?

d. The true mean and variance of X are $\mu = 1$ and $\sigma^2 = 1$. How close are the sample mean and the sample variance to the true values?

B.11 Use the recursive relationship in (B.52) with $X_0 = 1234567$, $m = 2^{32}$, $a = 1103515245$, and $c = 12345$ to generate 1,000 random values called *U1*. Do the resulting numbers appear random? Is this a good random number generator, or not? Choose another seed value and generate another 1,000 values called *U2*. Find the summary statistics and sample correlation for *U1* and *U2*. Do the values behave as you expect them to, or not?

B.12 Suppose that the joint *pdf* of the continuous random variables X and Y is $f(x, y) = 6x^2y$ for $0 \le x \le 1, 0 \le y \le 1$.

a. Does this function satisfy the conditions for a valid *pdf*?
b. Find the marginal *pdf* of X, as well as its mean and variance.
c. Find the marginal *pdf* of Y.
d. Find the conditional *pdf* of X given $Y = \frac{1}{2}$.
e. Find the conditional mean and variance of X given $Y = \frac{1}{2}$.
f. Are X and Y independent? Explain.

B.13 Suppose that X and Y are continuous random variables with joint *pdf* $f(x, y) = \frac{1}{2}$ for $0 \le x \le y \le 2$ and $f(x, y) = 0$ otherwise. Note that the values of X are less than or equal to the values of Y.

a. Verify that the volume under the joint *pdf* is 1.
b. Find the marginal *pdf*s of X and Y.
c. Find $P\left(X < \frac{1}{2}\right)$.
d. Find the *cdf* of Y.
e. Find the conditional probability $P\left(X < \frac{1}{2} \middle| Y = 1.5\right)$. Are X and Y independent?
f. Find the expected value and variance of Y.
g. Use the law of iterated expectations to find $E(X)$.

B.14 Let X and Y be two discrete random variables. X can take the values 1, 2, 3, or 4. Y can take the values 1, 2, 3. Their joint *pdf* is

		X			
		1	**2**	**3**	**4**
Y	1	0.01	0.07	0.09	0.03
	2	0.20	0	0.05	0.25
	3	0.09	0.03	0.06	0.12

a. Find the marginal distributions, the *pdf*s of X and Y.
b. Are these two random variables statistically independent? If not, give an example that disproves independence.
c. Find the conditional *pdf* of X given that $Y = 2$, $f(x|Y = 2)$, for $x = 1, 2, 3$, and 4.
d. Find the expected value of X.
e. Find the expected value of X given that $Y = 2$.

B.15 This exercise uses the random variables X and Y, and their joint *pdf*, from Exercise B.14.

a. Find the variance of X.
b. Find the variance of X given that $Y = 2$, and the variance of X given that $Y = 3$. Are they equal?
c. Find the conditional expectations $E(X|Y = 1)$, $E(X|Y = 2)$, and $E(X|Y = 3)$. Using these values show that $E(X) = \sum_{i=1}^{3} E(X|Y = i)\,P(Y = i)$.
d. Find $E(XY)$.
e. Find $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.
f. Find the correlation between X and Y.

B.16 Suppose that the two continuous random variables X and Y have joint *pdf* $f(x, y) = \frac{21}{4}x^2y$, if $x^2 \le y \le 1$.

a. Show that the marginal *pdf* of X is $f(x) = \int_{x^2}^{1} \frac{21}{4}x^2y\,dy = \frac{21}{8}x^2\left(1 - x^4\right)$ if $-1 \le x \le 1$.

b. Show that the conditional *pdf* of Y given that $X = \frac{1}{2}$ is $f(y|X = 1/2) = \frac{32}{15}y$ for $0.25 \le y \le 1$.

c. Show that the conditional *pdf* of Y given $X = x$ is $f(y|X = x) = \dfrac{2y}{1 - x^4}$ if $x^2 \le y \le 1$.

d. Using the result in part (c), show that $E(Y|X = x) = \left(\frac{2}{3}\right)\left(\dfrac{1 - x^6}{1 - x^4}\right)$.

e. It can be shown that the *pdf* of Y is $f(y) = \frac{7}{2}y^{5/2}$ if $0 \le y \le 1$. (i) Verify that this is a legitimate *pdf* and (ii) using this result show that $E(Y) = \frac{7}{9}$.

f. Using the results in (d) and (a), use the law of iterated expectations to show that $E(Y) = E_X[E(Y|X = x)] = \frac{7}{9}$.

B.17 Consider the random variable X which is the number of heads occurring in two flips of a fair coin.

a. What values can X take? What are the probabilities of each outcome? What is the probability that $X \le 1.5$?

b. Write down the values of the cumulative distribution function of X. What is the probability that $X \le 1$? What is the probability that $X \le 1.5$?

c. Suppose a wager is proposed in which you will receive winnings of $W = 2X$ dollars. What is the probability distribution of W?

d. What are your expected winnings? Show your work.

e. What is the conditional probability density function of X given that the first flip is a head?

f. What is the conditional expectation of $W = 2X$ given that the first flip is a head?

B.18 Suppose X is a continuous random variable that can take any value between zero and three, $0 < x < 3$. The *pdf* is $f(x) = cx^2$.

a. Find the value of c that makes this a legitimate *pdf*.

b. Using the result in (a) find $P(0 < X < 2)$. Show your work.

c. Find the mathematical equation for the *cdf* $F(x)$. Draw a sketch of the *cdf* for $-\infty < x < \infty$.

d. Use the *cdf* in (c) to compute $P(0.5 < X < 1)$.

e. Find the probability $P(0.5 < X < 1)$ given that $X < 2$.

B.19 The *cdf* of the continuous random variable X is $F(x) = 1 - e^{-2x}$ for $x \ge 0$ and $F(x) = 0$ otherwise.

a. Draw a sketch of the *cdf*.

b. Use the *cdf* to find the probability $P(1 < X < 2)$.

c. Find the *pdf* of X. Sketch the *pdf*.

d. Sketch on the *pdf* the area representing $P(1 < X < 2)$.

B.20 Two discrete random variables X and Y have the joint *pdf* $f(x, y) = c(2x + y)$. The random variable X takes the values $x = 0,\ 1,\ 2$ and the random variable Y takes the values $y = 0,\ 1,\ 2,\ 3$.

a. Find the value c that makes the probabilities sum to 1.

b. Find $P(X \ge 1, Y \le 1)$.

c. Find the marginal *pdf*s of X and Y.

d. Find the probability $P(X \ge 1, Y \le 1)$ given that $Y \le 2$.

e. Find the expected value of X.

f. Find the expected value of X given that $Y \le 2$.

g. Are X and Y statistically independent? Explain.

B.21 This exercise uses the joint *pdf* in Exercise B.14.

a. Find the variance of Y.

b. Find $E(Y|X = 1)$, $E(Y|X = 2)$, $E(Y|X = 3)$, and $E(Y|X = 4)$.

c. Calculate $\sum_{x=1}^{4}[E(Y|X = x) - E(Y)]^2 f(x)$. Which term in equation (B.27) does this represent?

d. Find $\text{var}(Y|X = 1)$, $\text{var}(Y|X = 2)$, $\text{var}(Y|X = 3)$, $\text{var}(Y|X = 4)$.

e. Calculate $\sum_{x=1}^{4}[\text{var}(Y|X = x)]\, f(x)$. Which term in equation (B.27) does this represent?

f. Use the results in parts (c) and (e) to compute $\text{var}(Y)$.

B.22 An econometrics instructor randomly chooses $n = 5$ students and gives each a problem to solve. Let the random variables $X_i = 1$ if the ith student answers correctly and $X_i = 0$ if the student does not answer the question correctly. Suppose that the probability that each student answers correctly is 0.80. Let $X = \sum_{i=1}^{5} X_i$ be the number of students who answer correctly.

a. Use the binomial distribution (B.43) to compute $P(X = 3|n = 5,\ p = 0.80)$.

b. From the first group of five students selected, four answered correctly. From a second randomly selected group of five students, two answered correctly. How does this illustrate the concept of sampling variation?

c. The instructor repeats the experiment of randomly selecting five students many times, recording the value X in each experiment. What will the average number of students answering correctly converge toward, as the number of experiments becomes very large?

d. Draw a sketch of the *pdf* of this random variable, locating $E(X)$ on the graph.

e. Find var(X). How does this value relate to the concept of sampling variation?

B.23 Suppose that for the population of married U.S. women, the average number of extramarital affairs, X, is $\mu = 2$.

a. Use the Poisson density function in (B.44) to find the probability that a randomly chosen married women will have $X = 2$ affairs.

b. Find the probability that a randomly chosen married woman will have two or more extramarital affairs. [*Hint:* First compute $P(X = 0)$ and $P(X = 1)$.]

c. Instead of sampling the entire population of married U.S. women, suppose that we sample the population of women who are known to have had at least one extramarital affair. Find the probability that a randomly chosen married woman will have two or more extramarital affairs given that she will have had at least one. That is, find $P(X \geq 2 | X \geq 1)$.

B.24 **Chebyshev's inequality** is a remarkable statistical result. Suppose X is a discrete or continuous random variable with mean μ and variance σ^2. Let ε be any positive number, then $P(|X - \mu| \geq \varepsilon) \leq \sigma^2/\varepsilon^2$.

a. Let X be a normal random variable with mean $\mu = 1$ and variance $\sigma^2 = 1$. Draw a sketch of the *pdf* of X.

b. Let $\varepsilon = 1$. On the sketch in (a) show $P(|X - 1| \geq 1)$.

c. Using the normal probabilities in Statistical Table 1, or your computer software, compute $P(|X - 1| \geq 1)$. Does the calculated value agree with Chebyshev's inequality?

B.25 Chebyshev's inequality is given in Exercise B.24.

a. If we let $\varepsilon = k\sigma$ what does the inequality become?

b. Let X be a normal random variable with mean $\mu = 1$ and variance $\sigma^2 = 1$. Find the exact probability $P(|X - 1| \geq 2\sigma)$. Does the value you calculate agree with the version of Chebyshev's inequality derived in part (a)?

c. Let U be a uniform random variable, see Section B.3.4, on the interval [0, 1]. Find the exact probability $P(|U - 0.5| > 2\sigma)$. Does this result agree with the revised version of Chebyshev's inequality derived in part (a)?

d. Let Y be a binomial random variable based on $n = 10$ trials each with probability $p = 0.8$. For this binomial distribution, what are the mean μ and standard deviation σ? Using your computer software, compute $P(|Y - \mu| > 2\sigma)$. Does your computed value agree with the revised version of Chebyshev's inequality derived in part (a)?

B.26 Suppose that X is a random variable, and $g(X)$ is a *convex* function of X. Then **Jensen's inequality**, as used in probability theory, says $g[E(X)] \leq E\left[g(X)\right]$. A convex function "curves up" without any inflection points. If a function $g(X)$ has second derivative that is positive over an interval, then it is convex over the interval.

a. Consider the function $g(X) = X^2$ over the interval $X > 0$. Find the second derivative of this function. Is $g(X)$ convex for $X > 0$? Draw a simple sketch of the function.

b. Suppose X is a discrete random variable taking the values $x = 1, 2, 3, 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Find $E(X)$ and $E(X^2)$. Is $[E(X)]^2 \leq E\left(X^2\right)$?

c. The variance of the random variable X is $E\left\{[X - E(X)]^2\right\} = E\left(X^2\right) - [E(X)]^2$. Using Jensen's inequality what can we say about the variance of a random variable?

B.27 Suppose that X is a random variable, and $g(X)$ is a *concave* function of X. Then Jensen's inequality, as used in probability theory, says $g[E(X)] \geq E\left[g(X)\right]$. A concave function has a continuously diminishing slope. If a function $g(X)$ has second derivative that is negative over an interval, then it is concave over the interval.

a. Consider the function $g(X) = \ln(X)$ over the interval $X > 0$. Find the second derivative of this function. Is $g(X)$ concave for $X > 0$? Draw a simple sketch of the function.

b. Suppose X is a discrete random variable taking the values $x = 1, 2, 3, 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Find $E(X)$ and $E[\ln(X)]$. Is $\ln[E(X)] \geq E[\ln(X)]$?

c. Jensen's inequality is also true for sample averages. Suppose $x_1, x_2, \ldots, x_n$ are numbers and $g(x)$ is a concave function. Then $g\left(\sum_{i=1}^{n} x_i/n\right) \geq \sum_{i=1}^{n} g(x_i)/n$. Suppose $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, and $x_4 = 4$. Show that $\ln\left(\sum_{i=1}^{4} x_i/4\right) \geq \sum_{i=1}^{4} \ln(x_i)/4$.

B.28 Let X and Y be random variables. The **Cauchy–Schwarz inequality**, as used in probability theory, is $[E(XY)]^2 \leq E(X^2)\,E(Y^2)$.

a. Using the joint probabilities in Table P.3, in the Probability Primer, Section P.3.2, verify that $[E(XY)]^2 \leq E(X^2)\,E(Y^2)$ holds.

b. Replace the random variables X and Y by $X - E(X) = X - \mu_X$ and $Y - E(Y) = Y - \mu_Y$. Show that the Cauchy–Schwarz inequality implies $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$.

c. Using the joint probabilities in Table P.3, in the Probability Primer, Section P.3.2, verify that $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$.

d. Use the fact that $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$ to prove that the correlation ρ_{XY} must fall in the interval $[-1, 1]$.

e. Show that $[\text{cov}(X, Y)]^2 = \text{var}(X)\,\text{var}(Y)$ if $Y = a + bX$, where a and b are constants.

B.29 Let X be a random variable and consider a function $g(X) \geq 0$ for every value of X. Assume $E[g(X)]$ exists. Then **Markov's inequality** is $P(g(X) \geq c) \leq c^{-1}E[g(X)]$.

a. Suppose X is a discrete random variable taking the values $x = 1,\ 2,\ 3,\ 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Let $g(X) = X^2$. Find $P[X^2 \geq 5]$. Find $E(X^2)$. Is $P[X^2 \geq 5] \leq E(X^2)/5$?

b. Let $g(X) = (X - \mu_X)^2$, where $\mu_X = E(X)$. Let $c = k^2\sigma_X^2$. Show that Markov's inequality leads to Chebyshev's inequality. [Author's note: Many mathematical inequalities are used in probability and statistics. A good list is in Dale J. Poirier (1995) *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Chapter 2.8. There Poirier (page 76) also relates a conversation between Nobel Prize winning economist Lawrence Klein and statistician Harold Freeman. *Lawrence Klein* "If the Devil promised you a theorem in return for your immortal soul, would you accept the bargain?" *Harold Freeman* "No. But I would for an inequality."]

B.30 Suppose X is a uniformly distributed variable on the (0, 1) interval. That is, $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. Further, suppose that the conditional *pdf* of Y given $X = x$ is $f(y|x) = 1/x$ for $0 < y < x$ and $f(y|x) = 0$ otherwise. [Adapted from Takeshi Amemiya (1994) *Introduction to Statistics and Econometrics*, Harvard University Press.]

a. Use the law of iterated expectations to show that $E(Y) = E_X E(Y|X) = 1/4$.

b. Show that $f(x, y) = 1/x$ for $0 < x < 1$ and $0 < y < x$, but $f(x, y) = 0$ otherwise. Then show $f(y) = \ln(y)$ for $0 < y < 1$. Then find $E(Y) = \displaystyle\int_0^1 y f(y)\,dy$.

B.31 Suppose X is a uniformly distributed variable on the (0, 1) interval. That is, $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. Suppose the random variable Y takes the values 1 and 0, and the conditional probabilities of these values are $P(Y = 1|X = x) = x$ and $P(Y = 0|X = x) = 1 - x$. [Adapted from Takeshi Amemiya (1994) *Introduction to Statistics and Econometrics*, Harvard University Press.]

a. Use the law of iterated expectations to show $E(Y) = 1/2$.

b. Use the variance decomposition to show that $\text{var}(Y) = 1/4$.

Appendix C

Review of Statistical Inference

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Discuss the difference between a population and a sample, and why we use samples of data as a basis for inference about population parameters.
2. Connect the concepts of a population and a random variable, indicating how the probability density function of a random variable, and the expected value and variance of the random variable, inform us about the population.
3. Explain the difference between the population mean and the sample mean.
4. Explain the difference between an estimate and an estimator, and why the latter is a random variable.
5. Explain the terms sampling variation and sampling distribution.
6. Explain the concept of unbiasedness, and use the rules of expected values to show that the sample mean is unbiased.
7. Explain why we prefer unbiased estimators with smaller variances to those with larger variances.
8. Describe the central limit theorem, and its implications for statistical inference.
9. Explain the relation between the population ''standard deviation'' and the standard error of the sample mean.
10. Explain the difference between point and interval estimation, and construct and interpret interval estimates of a population mean given a sample of data.
11. Give, in simple terms, a clarification of what the phrase ''95% level of confidence'' does and does not mean in relation to interval estimation.
12. Explain the purpose of hypothesis testing, and list the elements that must be present when carrying out a test.
13. Discuss the implications of the possible alternative hypotheses when testing the null hypothesis $H_0 : \mu = 7$. Give an economic example in which this hypothesis might be tested against one of the alternatives.
14. Describe the level of significance of a test, and explain the difference between the level of significance and the *p*-value of a test.
15. Define Type I error and its relationship to the level of significance of a test.
16. Explain the difference between one-tail tests and two-tail tests, describing when one is preferred to the other.
17. Explain the difference and implications between the statements ''I accept the null hypothesis'' and ''I do not reject the null hypothesis.''
18. Give an intuitive explanation of maximum likelihood estimation, and describe the properties of the maximum likelihood estimator.
19. List the three types of tests associated with maximum likelihood estimation and comment on their similarities and differences.
20. Distinguish between parametric and nonparametric estimation.
21. Understand how a kernel density estimator fits an empirical distribution.

KEYWORDS

alternative hypothesis
asymptotic distribution
BLUE
central limit theorem
central moments
estimate
estimator
experimental design
information measure
interval estimate
kernel density estimator
Lagrange multiplier test
law of large numbers
level of significance
likelihood function
likelihood ratio test
linear estimator
log-likelihood function
maximum likelihood estimation
nonparametric
null hypothesis
parametric
point estimate
population parameter
p-value
random sample
rejection region
sample mean
sample variance
sampling distribution
sampling variation
standard error
standard error of the estimate
standard error of the mean
statistical inference
test statistic
two-tail tests
Type I error
Type II error
unbiased estimators
Wald test

Economists are interested in relationships between economic variables. For example, how much can we expect the sales of Frozen Delight ice cream to rise if we reduce the price by 5%? How much will household food expenditure rise if household income rises by $100 per month? Questions such as these are the main focus of this book.

However, sometimes questions of interest focus on a single economic variable. For example, an airplane seat designer must consider the average hip size of passengers in order to allow adequate room for each person, while still designing the plane to carry the profit-maximizing number of passengers. What is the average hip size, or more precisely hip width, of U.S. flight passengers? If a seat 18 inches wide is planned, what percent of customers will not be able to fit? Questions like this must be faced by manufacturers of everything from golf carts to women's jeans. How can we answer these questions? We certainly cannot take the measurements of every man, woman, and child in the U.S. population. This is a situation when statistical inference is used. Infer means "to conclude by reasoning from something known or assumed." **Statistical inference** means that we will draw conclusions about a population based on a sample of data.

C.1 A Sample of Data

To carry out statistical inference, we need data. The data should be obtained from the population in which we are interested. For the airplane seat designer this is essentially the entire U.S. population above the age of two, since small children can fly "free" on the laps of their suffering parents. A separate branch of statistics, called **experimental design**, is concerned with the question of how to actually collect a representative sample. How would you proceed if you were asked to obtain 50 measurements of hip size representative of the entire population? This is not such an easy task. Ideally the 50 individuals will be randomly chosen from the population, in such a way that there is no pattern of choices. Suppose we focus on only the population of adult flyers, since usually there are few children on planes. Our experimental design specialist draws a sample that is shown in Table C.1 and stored in the data file *hip*.

TABLE C.1 **Sample Hip Size Data**

14.96	14.76	15.97	15.71	17.77
17.34	17.89	17.19	13.53	17.81
16.40	18.36	16.87	17.89	16.90
19.33	17.59	15.26	17.31	19.26
17.69	16.64	13.90	13.71	16.03
17.50	20.23	16.40	17.92	15.86
15.84	16.98	20.40	14.91	16.56
18.69	16.23	15.94	20.00	16.71
18.63	14.21	19.08	19.22	20.23
18.55	20.33	19.40	16.48	15.54

EXAMPLE C.1 | Histogram of Hip Width Data

A first step when analyzing a sample of data is to examine it visually. Figure C.1 is a histogram of the 50 data points. Based on this figure, the "average" hip size in this sample seems to be between 16 and 18 inches. For our profit-maximizing designer this casual estimate is not sufficiently precise. In the next section we set up an econometric model that will be used as a basis for inference in this problem.

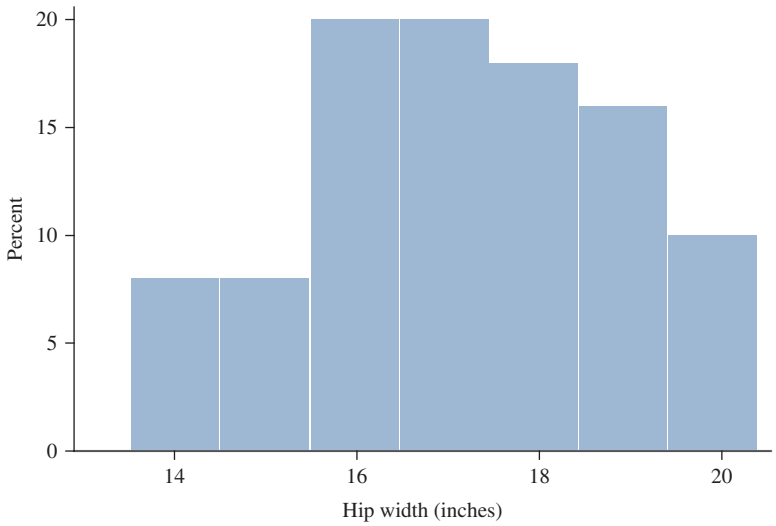


FIGURE C.1 **Histogram of hip sizes.**

C.2 An Econometric Model

The data in Table C.1 were obtained by sampling. Sampling from a population is an experiment. The variable of interest in this experiment is an individual's hip size. Before the experiment is performed we do not know what the values will be, thus the hip size of a randomly chosen person is a random variable. Let us denote this random variable as Y. We choose a sample of $N = 50$ individuals, $Y_1, Y_2, \ldots, Y_N$, where each Y_i represents the hip size of a different person. The data values in

Table C.1 are specific values of the variables, which we denote as $y_1, y_2, \ldots, y_N$. We assume that the population has a center, which we describe by the expected value of the random variable Y,

$$E(Y) = \mu \tag{C.1}$$

We use the Greek letter μ ("mu") to denote the mean of the random variable Y, and also the mean of the population we are studying. Thus if we knew μ we would have the answer to the question "What is the average hip size of adults in the United States?" To indicate its importance to us in describing the population we call μ a **population parameter**, or, more briefly, a parameter. Our objective is to use the sample of data in Table C.1 to make inferences, or judgments, about the unknown population parameter μ.

The other random variable characteristic of interest is its variability, which we measure by its variance,

$$\text{var}(Y) = E\left[Y - E(Y)\right]^2 = E\left[Y - \mu\right]^2 = \sigma^2 \tag{C.2}$$

The variance σ^2 is also an unknown population parameter. As described in the Probability Primer, the variance of a random variable measures the "spread" of a probability distribution about the population mean, with a larger variance meaning a wider spread, as shown in Figure P.3. In the context of the hip data, the variance tells us how much hip sizes can vary from one randomly chosen person to the next. To economize on space, we will denote the mean and variance of a random variable as $Y \sim (\mu, \sigma^2)$ where $\sim$ means "is distributed as." The first element in parentheses is the population mean and the second is the population variance. So far we have not said what kind of probability distribution we think Y has.

The econometric model is not complete. If our sample is drawn randomly, we can assume that $Y_1, Y_2, \ldots, Y_N$ are statistically independent. The hip size of any one individual is independent of the hip size of another randomly drawn individual. Furthermore, we assume that each of the observations we collect is from the population of interest, so each random variable Y_i, has the same mean and variance, or $Y_i \sim (\mu, \sigma^2)$. The Y_i constitute a **random sample**, in the statistical sense, because $Y_1, Y_2, \ldots, Y_N$ are statistically independent with identical probability distributions. It is sometimes reasonable to assume that population values are *normally* distributed, which we represent by $Y \sim N(\mu, \sigma^2)$.

C.3 Estimating the Mean of a Population

How shall we estimate the population mean μ given our sample of data values in Table C.1? The population mean is given by the expected value $E(Y) = \mu$. The expected value of a random variable is its average value in the population. It seems reasonable, by analogy, to use the average value in the sample, or **sample mean**, to estimate the population mean. Denote by $y_1, y_2, \ldots, y_N$ the sample of N observations. Then the sample mean is

$$\bar{y} = \sum y_i / N \tag{C.3}$$

The notation $\bar{y}$ (pronounced "y-bar") is widely used for the sample mean, and you probably encountered it in your statistics courses.

EXAMPLE C.2 | Sample Mean of Hip Width Data

For the hip data in Table C.1 we obtain $\bar{y} = 17.1582$, thus we estimate that the average hip size in the population is 17.1582 inches.

Given the estimate $\bar{y} = 17.1582$ we are inclined to ask, "How good an estimate is 17.1582?" By that we mean how close is 17.1582 to the true population mean, μ? Unfortunately this is an ill-posed question in the sense that it can never be answered. In order to answer it, we would have to know μ, in which case we would not have tried to estimate it in the first place!

Instead of asking about the quality of the *estimate* we will ask about the quality of the *estimation procedure*, or **estimator**. How good is the sample mean as an estimator of the mean of a population? This is a question we can answer. To distinguish between the estimate and the estimator of the population mean μ we will write the estimator as

$$\overline{Y} = \sum_{i=1}^{N} Y_i / N \tag{C.4}$$

In (C.4) we have used Y_i instead of y_i to indicate that this general formula is used whatever the sample values turn out to be. In this context Y_i are random variables, and thus the estimator $\overline{Y}$ is random too. We do not know the value of the estimator $\overline{Y}$ until a data sample is obtained, and different samples will lead to different values.

EXAMPLE C.3 | Sampling Variation of Sample Means of Hip Width Data

To illustrate, we collect 10 more samples of size $N = 50$ and calculate the average hip size, as shown in Table C.2. The estimates differ from sample to sample because $\overline{Y}$ is a random variable. This variation, due to collection of different random samples, is called **sampling variation**. It is an inescapable fact of statistical analysis that the estimator $\overline{Y}$—indeed, all statistical estimation procedures—are subject to sampling variability. Because of this terminology, an estimator's probability density function is called its **sampling distribution**.

TABLE C.2 Sample Means from 10 Samples

Sample	$\bar{y}$
1	17.3544
2	16.8220
3	17.4114
4	17.1654
5	16.9004
6	16.9956
7	16.8368
8	16.7534
9	17.0974
10	16.8770

We can determine how good the estimator $\overline{Y}$ is by examining its expected value, variance, and sampling distribution.

C.3.1 The Expected Value of $\overline{Y}$

Write out formula (C.4) fully as

$$\overline{Y} = \sum_{i=1}^{N} \frac{1}{N} Y_i = \frac{1}{N} Y_1 + \frac{1}{N} Y_2 + \cdots + \frac{1}{N} Y_N \tag{C.5}$$

From (P.16) the expected value of this sum is the sum of expected values

$$\begin{aligned} E\left(\overline{Y}\right) &= E\left(\frac{1}{N} Y_1\right) + E\left(\frac{1}{N} Y_2\right) + \cdots + E\left(\frac{1}{N} Y_N\right) \\ &= \frac{1}{N} E(Y_1) + \frac{1}{N} E(Y_2) + \cdots + \frac{1}{N} E(Y_N) \\ &= \frac{1}{N}\mu + \frac{1}{N}\mu + \cdots + \frac{1}{N}\mu \\ &= \mu \end{aligned}$$

The expected value of the estimator $\overline{Y}$ *is* the population mean μ that we are trying to estimate. What does this mean? The expectation of a random variable is its average value in all possible random samples from the population. If we did obtain many samples of size N, and obtained their average values, like those in Table C.2, then the average of all *those* values would equal the true population mean μ. This property is a good one for estimators to have. Estimators with this property are called **unbiased estimators**. The sample mean $\overline{Y}$ is an unbiased estimator of the population mean μ.

Unfortunately, while unbiasedness is a good property for an estimator to have, it does not tell us anything about whether our estimate $\bar{y} = 17.1582$, based on a single sample of data, is close to the true population mean value μ. To assess how far the estimate might be from μ, we will determine the variance of the estimator.

C.3.2 The Variance of $\overline{Y}$

The variance of $\overline{Y}$ is obtained using the procedure for finding the variance of a sum of uncorrelated (zero covariance) random variables in (P.23). We can apply this rule if our data are obtained by random sampling, because with random sampling the observations are statistically independent, and thus are uncorrelated. Furthermore, we have assumed that $\text{var}(Y_i) = \sigma^2$ for all observations. Carefully note how these assumptions are used in the derivation of the variance of $\overline{Y}$, which we write as $\text{var}(\overline{Y})$:

$$\begin{aligned}
\text{var}(\overline{Y}) &= \text{var}\left(\frac{1}{N}Y_1 + \frac{1}{N}Y_2 + \cdots + \frac{1}{N}Y_N\right) \\
&= \frac{1}{N^2}\text{var}(Y_1) + \frac{1}{N^2}\text{var}(Y_2) + \cdots + \frac{1}{N^2}\text{var}(Y_N) \\
&= \frac{1}{N^2}\sigma^2 + \frac{1}{N^2}\sigma^2 + \cdots + \frac{1}{N^2}\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned} \tag{C.6}$$

This result tells us that (i) the variance of $\overline{Y}$ is *smaller* than the population variance, because the sample size $N \geq 2$, and (ii) the larger the sample size, the smaller the sampling variation of $\overline{Y}$ as measured by its variance.

C.3.3 The Sampling Distribution of $\overline{Y}$

If the population data are normally distributed, then we say that the random variable Y_i follows a normal distribution. In this case the estimator $\overline{Y}$ also follows a normal distribution. In (P.36) it is noted that weighted averages of normal random variables are normal themselves. From (C.5) we know that $\overline{Y}$ is a weighted average of Y_i. If $Y_i \sim N(\mu, \sigma^2)$, then $\overline{Y}$ is also normally distributed, or $\overline{Y} \sim N(\mu, \sigma^2/N)$.

We can gain some intuition about the meaning and usefulness of the finding that $\overline{Y} \sim N(\mu, \sigma^2/N)$ if we examine Figure C.2. Each of the normal distributions in this figure is a sampling distribution of $\overline{Y}$. The differences among them are the sample sizes used in estimation. The sample size $N_3 > N_2 > N_1$. Increasing the sample size decreases the variance of the estimator $\overline{Y}$, $\text{var}(\overline{Y}) = \sigma^2/N$, and this increases the probability that the sample mean will be "close" to the true population parameter μ. When examining Figure C.2, recall that an area under a probability density function (*pdf*) measures the probability of an event. If ε represents a positive number, the probability that $\overline{Y}$ falls in the interval between $\mu - \varepsilon$ and $\mu + \varepsilon$ is greater for larger samples.

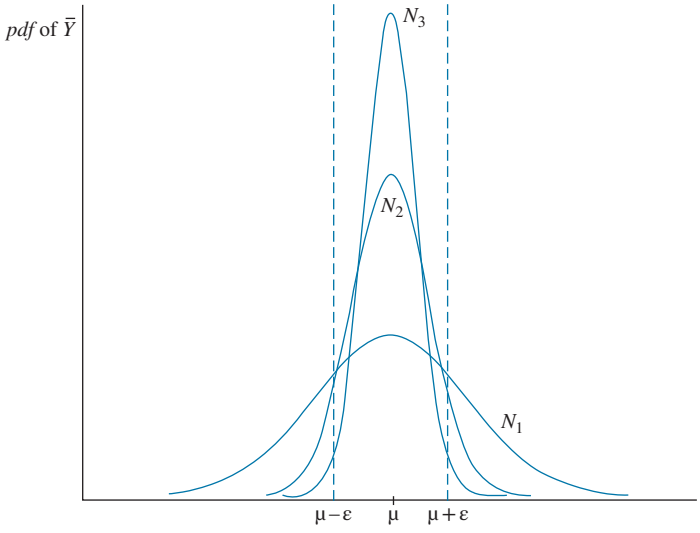


FIGURE C.2 **Increasing sample size and sampling distributions of $\bar{Y}$.**

The lesson here is that having more data is better than having less data, because having a larger sample increases the probability of obtaining an estimate "close" or "within ε" of the true population parameter μ.

EXAMPLE C.4 | The Effect of Sample Size on Sample Mean Precision

In our numerical example, suppose we want our estimate of μ to be within 1 inch of the true value. Let us compute the probability of getting an estimate within $\varepsilon = 1$ inch of μ—that is, within the interval $[\mu - 1, \mu + 1]$. For the purpose of illustration assume that the population is normal, $\sigma^2 = 10$ and $N = 40$. Then $\bar{Y} \sim N(\mu, \sigma^2/N = 10/40 = 0.25)$. We can compute the probability that $\bar{Y}$ is within 1 inch of μ by calculating $P[\mu - 1 \le \bar{Y} \le \mu + 1]$. To do so we standardize $\bar{Y}$ by subtracting its mean μ and dividing by its standard deviation $\sigma/\sqrt{N}$, and then use the standard normal distribution and Statistical Table 1:

$$P\left[\mu - 1 \le \bar{Y} \le \mu + 1\right] = P\left[\frac{-1}{\sigma/\sqrt{N}} \le \frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \le \frac{1}{\sigma/\sqrt{N}}\right]$$
$$= P\left[\frac{-1}{\sqrt{0.25}} \le Z \le \frac{1}{\sqrt{0.25}}\right]$$
$$= P[-2 \le Z \le 2] = 0.9544$$

Thus, if we draw a random sample of size $N = 40$ from a normal population with variance 10, using the sample mean as an estimator will provide an estimate within 1 inch of the true value about 95% of the time. If $N = 80$, the probability that $\bar{Y}$ is within 1 inch of μ increases to 0.995.

C.3.4 The Central Limit Theorem

We were able to carry out the above analysis because we assumed that the population we are considering, hip width of U.S. adults, has a normal distribution. This implies that $Y_i \sim N(\mu, \sigma^2)$, and $\bar{Y} \sim N(\mu, \sigma^2/N)$. A question we need to ask is "If the population is not normal, then what is the sampling distribution of the sample mean?" The **central limit theorem** provides an answer to this question.

Central Limit Theorem:

If $Y_1, \ldots, Y_N$ are independent and identically distributed random variables with mean μ and variance σ^2, and $\overline{Y} = \sum Y_i/N$, then

$$Z_N = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}}$$

has a probability distribution that converges to the standard normal $N(0, 1)$ as $N \to \infty$.

This theorem says that the sample average of N independent random variables from *any* probability distribution will have an approximate standard normal distribution after standardizing (i.e., subtracting the mean and dividing by the standard deviation), if the sample is sufficiently large. A shorthand notation is $\overline{Y} \overset{a}{\sim} N(\mu, \sigma^2/N)$, where the symbol $\overset{a}{\sim}$ means *asymptotically distributed*. The word **asymptotic** implies that the approximate normality of $\overline{Y}$ depends on having a large sample. Thus even if the population is not normal, if we have a sufficiently large sample, we can carry out calculations like those in the previous section. How large does the sample have to be? In general, it depends on the complexity of the problem, but in the simple case of estimating a population mean, if $N \geq 30$ then you can feel pretty comfortable in assuming that the sample mean is approximately normally distributed, $\overline{Y} \overset{a}{\sim} N(\mu, \sigma^2/N)$, as indicated by the central limit theorem.

EXAMPLE C.5 | Illustrating the Central Limit Theorem

To illustrate how well the central limit theorem actually works, we carry out a simulation experiment. Let the continuous random variable Y have a triangular distribution, with probability density function

$$f(y) = \begin{cases} 2y & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

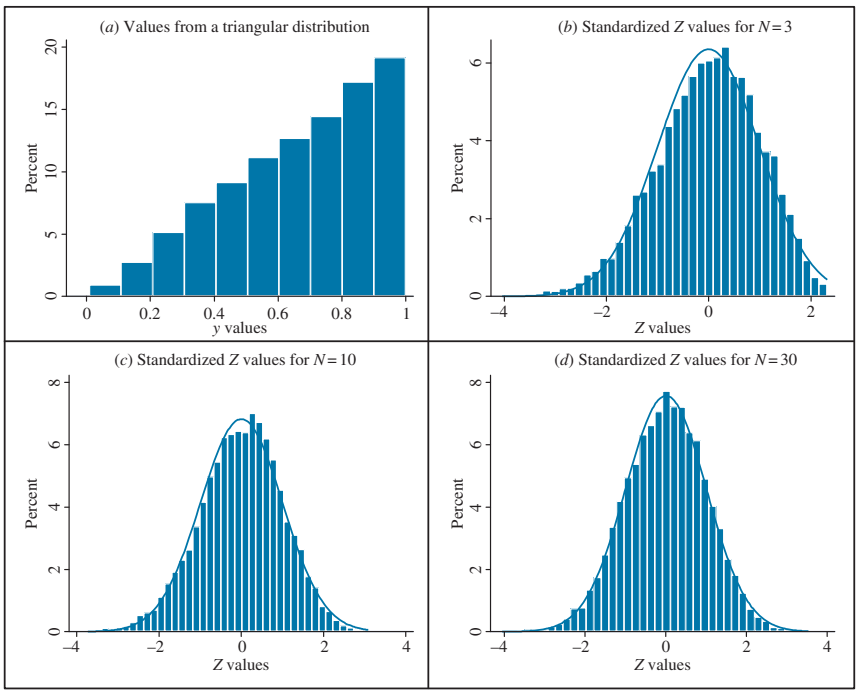


FIGURE C.3 **Central limit theorem.**

Draw a sketch of the triangular *pdf* to understand its name. The expected value of Y is $\mu = E(Y) = 2/3$, and its variance is $\sigma^2 = \text{var}(Y) = 1/18$. The central limit theorem says that if $Y_1, \ldots, Y_N$ are independent and identically distributed with density $f(y)$ then

$$Z_N = \frac{\overline{Y} - 2/3}{\sqrt{\frac{1/18}{N}}}$$

has a probability distribution that approaches the standard normal distribution as N approaches infinity.

We use a random number generator to create random values from the triangular *pdf*. Plotting 10,000 values gives the histogram in Figure C.3(a). We generate 10,000 samples of sizes $N = 3$, 10, and 30, compute the sample means of each sample, and create Z_N. Their histograms are shown in Figures C.3(b)–(d). You see the amazing convergence of the standardized sample mean's distribution to a distribution that is bell shaped, centered at zero, symmetric, with almost all values between -3 and 3, just like a standard normal distribution, with a sample size as small as $N = 10$.

C.3.5 Best Linear Unbiased Estimation

Another powerful finding about the estimator $\overline{Y}$ of the population mean is that it is the best of all possible estimators that are both *linear* and *unbiased*. A **linear estimator** is simply one that is a weighted average of Y_i's, such as $\overline{Y} = \sum a_i Y_i$, where a_i are constants. The sample mean $\overline{Y}$, given in (C.4), is a linear estimator with $a_i = 1/N$. The fact that $\overline{Y}$ is the "best" linear unbiased estimator (**BLUE**) accounts for its wide use. "Best" means that it is the linear unbiased estimator with the smallest possible variance. In the previous section we demonstrated that it is better to have an estimator with a smaller variance rather than a larger one—because it increases the chances of getting an estimate close to the true population mean μ. This important result about the estimator $\overline{Y}$ is true *if* the sample values $Y_i \sim (\mu, \sigma^2)$ are uncorrelated and identically distributed. It does not depend on the population being normally distributed. A proof of this result is in Section C.9.2.

C.4 Estimating the Population Variance and Other Moments

The sample mean $\overline{Y}$ is an estimate of the population mean μ. The population mean is often called the "first moment" since it is the expected value of Y to the first power. Higher moments are obtained by taking expected values of higher powers of the random variable, so the second moment of Y is $E(Y^2)$, the third moment is $E(Y^3)$, and so on. When the random variable has its population mean subtracted, it is said to be *centered*. Expected values of powers of centered random variables are called **central moments**, and they are often denoted as μ_r, so that the rth central moment of Y is

$$\mu_r = E[(Y - \mu)^r]$$

The value of the first central moment is zero since $\mu_1 = E(Y - \mu)^1 = E(Y) - \mu = 0$. It is the higher central moments of Y that are interesting:

$$\mu_2 = E[(Y - \mu)^2] = \sigma^2$$
$$\mu_3 = E[(Y - \mu)^3]$$
$$\mu_4 = E[(Y - \mu)^4]$$

You recognize that the second central moment of Y is its variance, and the third and fourth moments appear in the definitions of skewness and kurtosis introduced in Appendix B.1.2. The question we address in this section is, now that we have an excellent estimator of the mean of a population, how do we estimate these higher moments? We will first consider estimation of the population variance, and then address the problem of estimating the third and fourth moments.

C.4.1 Estimating the Population Variance

The population variance is $\text{var}(Y) = \sigma^2 = E[Y - \mu]^2$. An expected value is an "average" of sorts, so if we knew μ we could estimate the variance by using the sample analog $\tilde{\sigma}^2 = \sum(Y_i - \mu)^2/N$. We do not know μ, so replace it by its estimator $\overline{Y}$, giving

$$\tilde{\sigma}^2 = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{N}$$

This estimator is not a bad one. It has a logical appeal, and it can be shown to converge to the true value of σ^2 as the sample size $N \to \infty$, but it is biased. To make it unbiased, we divide by $N - 1$ instead of N. This correction is needed since the population mean μ has to be estimated before the variance can be estimated. This change does not matter much in samples of at least 30 observations, but it does make a difference in smaller samples. The unbiased estimator of the population variance σ^2 is

$$\hat{\sigma}^2 = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{N - 1} \tag{C.7}$$

You may remember this estimator from a prior statistics course as the "sample variance." Using the **sample variance** we can estimate the variance of the estimator $\overline{Y}$ as

$$\widehat{\text{var}}\left(\overline{Y}\right) = \hat{\sigma}^2/N \tag{C.8}$$

In (C.8) note that we have put a "hat" ($\hat{}$) over this variance to indicate that it is an estimated variance. The square root of the estimated variance is called the **standard error** of $\overline{Y}$ and is also known as the **standard error of the mean** and the **standard error of the estimate**,

$$\text{se}\left(\overline{Y}\right) = \sqrt{\widehat{\text{var}}\left(\overline{Y}\right)} = \hat{\sigma}/\sqrt{N} \tag{C.9}$$

C.4.2 Estimating Higher Moments

Recall that central moments are expected values, $\mu_r = E\left[(Y - \mu)^r\right]$, and thus are averages in the population. In statistics the **law of large numbers** says that sample means converge to population averages (expected values) as the sample size $N \to \infty$. We can estimate the higher moments by finding the sample analog and replacing the population mean μ by its estimate $\overline{Y}$, so that

$$\tilde{\mu}_2 = \sum\left(Y_i - \overline{Y}\right)^2/N = \tilde{\sigma}^2$$

$$\tilde{\mu}_3 = \sum\left(Y_i - \overline{Y}\right)^3/N$$

$$\tilde{\mu}_4 = \sum\left(Y_i - \overline{Y}\right)^4/N$$

Note that in these calculations we divide by N and not by $N - 1$, since we are using the law of large numbers (i.e., large samples) as justification, and in large samples the correction has little effect. Using these sample estimates of the central moments we can obtain estimates of the skewness coefficient (S) and kurtosis coefficient (K) as

$$\widehat{skewness} = S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{kurtosis} = K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

EXAMPLE C.6 | Sample Moments of the Hip Data

The sample variance for the hip data is

$$\hat{\sigma}^2 = \frac{\sum(y_i - \bar{y})^2}{N-1} = \frac{\sum(y_i - 17.1582)^2}{49} = \frac{159.9995}{49} = 3.2653$$

This means that the estimated variance of the sample mean is

$$\widehat{\text{var}}(\bar{Y}) = \frac{\hat{\sigma}^2}{N} = \frac{3.2653}{50} = 0.0653$$

and the standard error of the mean is

$$\text{se}(\bar{Y}) = \hat{\sigma}/\sqrt{N} = 0.2556$$

The estimated skewness is $S = -0.0138$ and the estimated kurtosis is $K = 2.3315$ using

$$\tilde{\sigma} = \sqrt{\sum(Y_i - \bar{Y})^2/N} = \sqrt{159.9995/50} = 1.7889$$

$$\tilde{\mu}_3 = \sum(Y_i - \bar{Y})^3/N = -0.0791$$

$$\tilde{\mu}_4 = \sum(Y_i - \bar{Y})^4/N = 23.8748$$

Thus, the hip data is slightly negatively skewed and is slightly less peaked than would be expected for a normal distribution. Nevertheless, as we will see in Section C.7.4, we cannot conclude that the hip data follow a non-normal distribution.

EXAMPLE C.7 | Using the Hip Data Estimates

How can we summarize what we have learned? Our estimates suggest that the hip size of U.S. adults is normally distributed with mean 17.158 inches and with a variance of 3.265; $Y \sim N(17.158, 3.265)$. Based on this information, if an airplane seat is 18 inches wide, what percentage of customers will not be able to fit? We can recast this question as asking what the probability is that a randomly drawn person will have hips larger than 18 inches,

$$P(Y > 18) = P\left(\frac{Y-\mu}{\sigma} > \frac{18-\mu}{\sigma}\right)$$

We can give an approximate answer to this question by replacing the unknown parameters by their estimates,

$$\widehat{P(Y > 18)} \cong P\left(\frac{Y-\bar{y}}{\hat{\sigma}} > \frac{18-17.158}{1.8070}\right) = P(Z > 0.4659) = 0.3207$$

Based on our estimates, 32% of the population would not be able to fit into a seat that is 18 inches wide.

How large would a seat have to be to fit 95% of the population? If we let y^* denote the required seat size, then

$$\widehat{P(Y \le y^*)} \cong P\left(\frac{Y-\bar{y}}{\hat{\sigma}} \le \frac{y^* - 17.1582}{1.8070}\right) = P\left(Z \le \frac{y^* - 17.1582}{1.8070}\right) = 0.95$$

Using your computer software, or the table of normal probabilities, the value of Z such that $P(Z \le z^*) = 0.95$ is $z^* = 1.645$. Then

$$\frac{y^* - 17.1582}{1.8070} = 1.645 \Rightarrow y^* = 20.1305$$

Thus, to accommodate 95% of U.S. adult passengers, we estimate that the seats should be slightly greater than 20 inches wide.

C.5 Interval Estimation

In contrast to a **point estimate** of the population mean μ, like $\bar{y} = 17.158$, a confidence interval, or **interval estimate**, is a range of values that may contain the true population mean. A confidence interval contains information not only about the location of the population mean, but also about the precision with which we estimate it.

C.5.1 Interval Estimation: σ^2 Known

Let Y be a normally distributed random variable, $Y \sim N(\mu, \sigma^2)$. Assume that we have a random sample of size N from this population, $Y_1, Y_2, \ldots, Y_N$. The estimator of the population mean

is $\overline{Y} = \sum_{i=1}^{N} Y_i/N$. Because we have assumed that Y is normally distributed, it is also true that $\overline{Y} \sim N\left(\mu, \sigma^2/N\right)$.

For the present, let us assume that the population variance σ^2 is known. This assumption is not likely to be true, but making it allows us to introduce the notion of confidence intervals with few complications. In the next section we introduce methods for the case when σ^2 is unknown. Create a standard normal random variable

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/N}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1) \tag{C.10}$$

Cumulative probabilities for the standard normal are given by its cumulative distribution function (see the Probability Primer, Section P.7)

$$P(Z \leq z) = \Phi(z)$$

These values are given in Statistical Table 1. Let z_c be a "critical value" for the standard normal distribution, such that $\alpha = 0.05$ of the probability is in the tails of the distribution, with $\alpha/2 = 0.025$ of the probability in the tail to the right of z_c and $\alpha/2 = 0.025$ of the probability in the tail to the left of $-z_c$. The critical value is the 97.5 percentile of the standard normal distribution, $z_c = 1.96$, with $\Phi(1.96) = 0.975$. It is shown in Figure C.4. Thus, $P(Z \geq 1.96) = P(Z \leq -1.96) = 0.025$ and

$$P(-1.96 \leq Z \leq 1.96) = 1 - 0.05 = 0.95 \tag{C.11}$$

Substitute (C.10) into (C.11) and rearrange to obtain

$$P\left(\overline{Y} - 1.96\sigma/\sqrt{N} \leq \mu \leq \overline{Y} + 1.96\sigma/\sqrt{N}\right) = 0.95$$

In general,

$$P\left(\overline{Y} - z_c\frac{\sigma}{\sqrt{N}} \leq \mu \leq \overline{Y} + z_c\frac{\sigma}{\sqrt{N}}\right) = 1 - \alpha \tag{C.12}$$

where z_c is the appropriate critical value for a given value of tail probability α such that $\Phi\left(z_c\right) = 1 - \alpha/2$. In (C.12) we have defined the **interval estimator**

$$\overline{Y} \pm z_c\frac{\sigma}{\sqrt{N}} \tag{C.13}$$

Our choice of the phrase interval *estimator* is a careful one. Intervals constructed using (C.13), in repeated sampling from the population, have a $100(1 - \alpha)\%$ chance of containing the population mean μ.

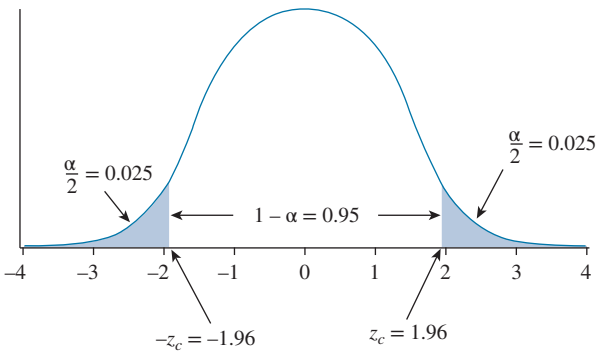


FIGURE C.4 **$\alpha = 0.05$ Critical values for the $N(0, 1)$ distribution.**

EXAMPLE C.8 | Simulating the Hip Data: Interval Estimates

In order to use the interval estimator in (C.13) we must have data from a normal population with a known variance. To illustrate the computation, and the meaning of interval estimation, we will create a sample of data using a computer simulation. Statistical software programs contain random number generators. These are routines that create values from a given probability distribution. Table C.3 (data file *table_c3*) contains 30 random values from a normal population with mean $\mu = 10$ and variance $\sigma^2 = 10$.

TABLE C.3 30 Values from $N(10, 10)$

11.939	11.407	13.809
10.706	12.157	7.443
6.644	10.829	8.855
13.187	12.368	9.461
8.433	10.052	2.439
9.210	5.036	5.527
7.961	14.799	9.921
14.921	10.478	11.814
6.223	13.859	13.403
10.123	12.355	10.819

The sample mean of these values is $\bar{y} = 10.206$ and the corresponding interval estimate for μ, obtained by applying the interval estimator in (C.13) with a 0.95 probability content, is $10.206 \pm 1.96 \times \sqrt{10/30} = [9.074, 11.338]$. To appreciate how the sampling variability of an interval estimator arises, consider Table C.4, which contains the interval estimate for the sample in Table C.3, as well as the sample means and interval estimates from another 9 samples of size 30, like that in Table C.3. The whole 10 samples are stored in the data file *table_c4*.

TABLE C.4 Confidence Interval Estimates from 10 Samples of Data

Sample	$\bar{y}$	Lower Bound	Upper Bound
1	10.206	9.074	11.338
2	9.828	8.696	10.959
3	11.194	10.063	12.326
4	8.822	7.690	9.953
5	10.434	9.303	11.566
6	8.855	7.723	9.986
7	10.511	9.380	11.643
8	9.212	8.080	10.343
9	10.464	9.333	11.596
10	10.142	9.010	11.273

Table C.4 illustrates the sampling variation of the estimator $\overline{Y}$. The sample mean varies from sample to sample. In this simulation, or Monte Carlo experiment, we know that the true population mean, $\mu = 10$, and the estimates $\overline{Y}$ are centered at that value. The half-width of the interval estimates is $1.96\sigma/\sqrt{N}$. Note that while the point estimates $\overline{Y}$ in Table C.4 fall near the true value $\mu = 10$, not all of the interval estimates contain the true value. Intervals from samples 3, 4, and 6 do not contain the true value $\mu = 10$. However, in 10,000 simulated samples the average value of $\bar{y} = 10.004$ and 94.86% of intervals constructed using (C.13) contain the true parameter value $\mu = 10$.

These numbers in Example C.8 reveal what is, and what is not, true about interval estimates.

- Any one interval estimate may or may not contain the true population parameter value.
- If *many* samples of size N are obtained, and intervals are constructed using (C.13) with $(1 - \alpha) = 0.95$, then 95% of them will contain the true parameter value.
- A 95% level of "confidence" is the probability that the interval estimator will provide an interval containing the true parameter value. Our confidence is in the procedure, not in any one interval estimate.

Since 95% of intervals constructed using (C.13) will contain the true parameter $\mu = 10$, we will be surprised if an interval estimate based on one sample does not contain the true parameter. Indeed, the fact that 3 of the 10 intervals in Table C.4 do not contain $\mu = 10$ is surprising, since out of 10 we would assume that only one 95% interval estimate might not contain the true parameter. This just goes to show that what happens in any one sample, or just a few samples, is not what sampling properties tell us. Sampling properties tell us what happens in many repeated experimental trials, or in all possible samples from a population.

C.5.2 Interval Estimation: σ^2 Unknown

The standardization in (C.10) assumes that the population variance σ^2 is known. When σ^2 is unknown, it is natural to replace it with its estimator $\hat{\sigma}^2$ given in (C.7)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}{N-1}$$

When we do so, the resulting standardized random variable has a t-distribution (see Appendix B.3.7) with $(N-1)$ degrees of freedom,

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)} \tag{C.14}$$

The notation $t_{(N-1)}$ denotes a t-distribution with $N-1$ "degrees of freedom." Let the critical value t_c be the $100(1-\alpha/2)$ -percentile value $t_{(1-\alpha/2,N-1)}$. This critical value has the property that $P\left[t_{(N-1)} \le t_{(1-\alpha/2,N-1)}\right] = 1-\alpha/2$. Critical values for the t-distribution are contained in Statistical Table 2. If t_c is a critical value from the t-distribution, then

$$P\left(-t_c \le \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \le t_c\right) = 1-\alpha$$

Rearranging, we obtain

$$P\left(\overline{Y} - t_c\frac{\hat{\sigma}}{\sqrt{N}} \le \mu \le \overline{Y} + t_c\frac{\hat{\sigma}}{\sqrt{N}}\right) = 1-\alpha$$

The $100(1-\alpha)\%$ interval estimator for μ is

$$\overline{Y} \pm t_c\frac{\hat{\sigma}}{\sqrt{N}} \quad \text{or} \quad \overline{Y} \pm t_c \text{se}\left(\overline{Y}\right) \tag{C.15}$$

Unlike the interval estimator for the known σ^2 case in (C.13), the interval in (C.15) has center *and width* that vary from sample to sample.

Remark

The confidence interval (C.15) is based upon the assumption that the population is normally distributed, so that $\overline{Y}$ is normally distributed. If the population is not normal, then we invoke the central limit theorem, and say that $\overline{Y}$ is approximately normal in "large" samples, which from Figure C.3 you can see might be as few as 30 observations. In this case, we can use (C.15), recognizing that there is an approximation error introduced in smaller samples.

EXAMPLE C.9 | Simulating the Hip Data: Continued

Table C.5 contains estimated values of σ^2 and interval estimates using (C.15) for the same 10 samples used for Table C.4. For the sample size $N = 30$ and the 95% confidence level, the t-distribution critical value $t_c = t_{(0.975,29)} = 2.045$. The estimates $\overline{Y}$ are the same as in Table C.4. The estimates $\hat{\sigma}^2$ vary about the true value $\sigma^2 = 10$. Of these 10 intervals, those for samples 4 and 6 do not contain the true parameter $\mu = 10$. Nevertheless, in 10,000 simulated samples 94.82% of them contain the true population mean $\mu = 10$.

TABLE C.5 **Interval Estimates Using (C.15) from 10 Samples**

Sample	$\bar{y}$	$\hat{\sigma}^2$	Lower Bound	Upper Bound
1	10.206	9.199	9.073	11.338
2	9.828	6.876	8.849	10.807
3	11.194	10.330	9.994	12.394
4	8.822	9.867	7.649	9.995
5	10.434	7.985	9.379	11.489
6	8.855	6.230	7.923	9.787
7	10.511	7.333	9.500	11.523
8	9.212	14.687	7.781	10.643
9	10.464	10.414	9.259	11.669
10	10.142	17.689	8.571	11.712

EXAMPLE C.10 | Interval Estimation Using the Hip Data

We have introduced the empirical problem faced by an airplane seat design engineer. Given a random sample of size $N = 50$ we estimated the mean U.S. hip width to be $\bar{y} = 17.158$ inches. Furthermore we estimated the population variance to be $\hat{\sigma}^2 = 3.265$; thus the estimated standard deviation is $\hat{\sigma} = 1.807$. The standard error of the mean is $\hat{\sigma}/\sqrt{N} = 1.807/\sqrt{50} = 0.2556$. The critical value for interval estimation comes from a t-distribution with $N - 1 = 49$ degrees of freedom. While this value is not in Statistical Table 2, the correct value using our software is $t_c = t_{(0.975,49)} = 2.0095752$, which we round to $t_c = 2.01$. To construct a 95% interval estimate we use (C.15), replacing estimates for the estimators, to give

$$\bar{y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} = 17.1582 \pm 2.01 \frac{1.807}{\sqrt{50}}$$
$$= [16.6447, 17.6717]$$

We estimate that the population mean hip size falls between 16.645 and 17.672 inches. Although we do not know if this interval contains the true population mean hip size for sure, we know that the procedure used to create the interval "works" 95% of the time; thus we would be surprised if the interval did not contain the true population value μ.

C.6 Hypothesis Tests About a Population Mean

Hypothesis testing procedures compare a conjecture, or a hypothesis, that we have about a population to the information contained in a sample of data. The conjectures we test here concern the mean of a normal population. In the context of the problem faced by the airplane seat designer, suppose that airplanes since 1970 have been designed assuming that the mean population hip width is 16.5 inches. Is that figure still valid today?

C.6.1 Components of Hypothesis Tests

Hypothesis tests use sample information about a parameter—namely, its point estimate and its standard error—to draw a conclusion about the hypothesis. In every hypothesis test, five ingredients must be present:

Components of Hypothesis Tests

A *null* hypothesis, H_0
An *alternative* hypothesis, H_1
A test *statistic*
A *rejection* region
A conclusion

The Null Hypothesis

The "null" hypothesis, which is denoted by H_0 (*H-naught*), specifies a value c for a parameter. We write the **null hypothesis** as $H_0: \mu = c$. A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true, in which case we *reject* the null hypothesis.

The Alternative Hypothesis

Paired with every null hypothesis is a logical alternative hypothesis, H_1, that we will accept if the null hypothesis is rejected. The **alternative hypothesis** is flexible and depends to some extent on the problem at hand. For the null hypothesis $H_0: \mu = c$ three possible alternative hypotheses are

- $H_1: \mu > c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ is greater than c.
- $H_1: \mu < c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ is less than c.
- $H_1: \mu \neq c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ takes a value other than (not equal to) c.

The Test Statistic

The sample information about the null hypothesis is embodied in the sample value of a **test statistic**. Based on the value of a test statistic, we decide either to reject the null hypothesis or not to reject it. A test statistic has a very special characteristic: its probability distribution is completely known when the null hypothesis is true, and it has some other distribution if the null hypothesis is not true.

Consider the null hypothesis $H_0: \mu = c$. If the sample data come from a normal population with mean μ and variance σ^2, then

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)}$$

If the null hypothesis $H_0: \mu = c$ is true, then

$$t = \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)} \tag{C.16}$$

If the null hypothesis is not true, then the t-statistic in (C.16) does not have the usual t-distribution.

Remark

The test statistic distribution in (C.16) is based on an assumption that the population is normally distributed. If the population is not normal, then we invoke the central limit theorem, and say that $\overline{Y}$ is approximately normal in "large" samples. We can use (C.16), recognizing that there is an approximation error introduced if our sample is small.

The Rejection Region The **rejection region** depends on the form of the alternative. It is the range of values of the test statistic that leads to rejection of the null hypothesis. They are values that are *unlikely* and have low probability of occurring when the null hypothesis is true. The chain of logic is "If a value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus it is unlikely that the null hypothesis is true." If the alternative hypothesis is true, then values of the test statistic will tend to be unusually large or unusually small. The terms "large" and "small" are determined by choosing a probability α, called the **level of significance** of the test, which provides a meaning for "an *unlikely* event." The level of significance of the test α is usually chosen to be 0.01, 0.05, or 0.10.

Conclusion When you have completed a hypothesis test, you should state your conclusion, whether you reject the null hypothesis. However, we urge you to make it standard practice to say what the conclusion means in the economic context of the problem you are working on—that is, interpret the results in a meaningful way. This should be a point of emphasis in all statistical work that you do.

We will now discuss the mechanics of carrying out alternative versions of hypothesis tests.

C.6.2 One-Tail Tests with Alternative "Greater Than" (>)

If the alternative hypothesis $H_1: \mu > c$ is true, then the value of the t-statistic (C.16) tends to become larger than usual for the t-distribution. Let the critical value t_c be the $100(1-\alpha)$-percentile $t_{(1-\alpha, N-1)}$ from a t-distribution with $N-1$ degrees of freedom. Then $P(t \le t_c) = 1-\alpha$, where α is the level of significance of the test. If the t-statistic is greater than or equal to t_c, then we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu > c$, as shown in Figure C.5.

If the null hypothesis $H_0: \mu = c$ is *true*, then the test statistic (C.16) has a t-distribution, and its values would tend to fall in the center of the distribution, where most of the probability is contained. If $t < t_c$, then there is no evidence against the null hypothesis, and we do not reject it.

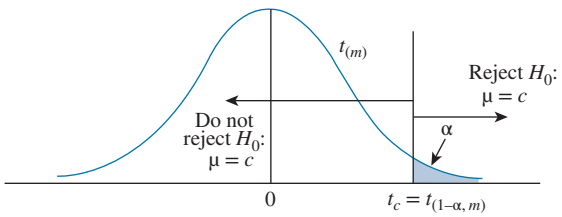


FIGURE C.5 **The rejection region for the one-tail test of $H_0: \mu = c$ against $H_1: \mu > c$.**

C.6.3 One-Tail Tests with Alternative "Less Than" (<)

If the alternative hypothesis $H_1: \mu < c$ is true, then the value of the t-statistic (C.16) tends to become smaller than usual for the t-distribution. The critical value $-t_c$ is the 100α-percentile $t_{(\alpha, N-1)}$ from a t-distribution with $N-1$ degrees of freedom. Then $P(t \le -t_c) = \alpha$, where α is the level of significance of the test as shown in Figure C.6. If $t \le -t_c$, then we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu < c$. If $t > -t_c$, then we do not reject $H_0: \mu = c$.

Memory Trick

The rejection region for a one-tail test is in the direction of the arrow in the alternative. If alternative is ">", then reject in right tail. If alternative is "<", reject in left tail.

C.6.4 Two-Tail Tests with Alternative "Not Equal To" (≠)

If the alternative hypothesis $H_1: \mu \neq c$ is true, then values of the test statistic may be unusually "large" or unusually "small." The rejection region consists of the two "tails" of the t-distribution, and this is called a **two-tail test**. In Figure C.7, the critical values for testing $H_0: \mu = c$ against $H_1: \mu \neq c$ are depicted. The critical value is the $100(1-\alpha/2)$-percentile from a t-distribution with $N-1$ degrees of freedom, $t_c = t_{(1-\alpha/2, N-1)}$, so that $P(t \ge t_c) = P(t \le -t_c) = \alpha/2$.

If the value of the test statistic t falls in the rejection region, either tail of the $t_{(N-1)}$ distribution, then we reject the null hypothesis $H_0: \mu = c$ and accept the alternative $H_1: \mu \neq c$. If the value of the test statistic t falls in the nonrejection region, between the critical values $-t_c$ and t_c, then we do not reject the null hypothesis $H_0: \mu = c$.

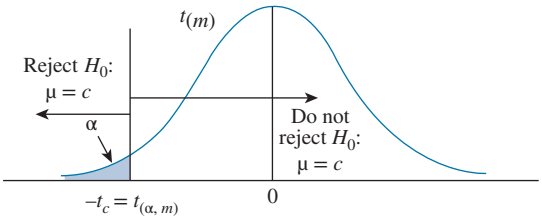


FIGURE C.6 **Critical value for one-tail test $H_0: \mu = c$ versus $H_1: \mu < c$.**

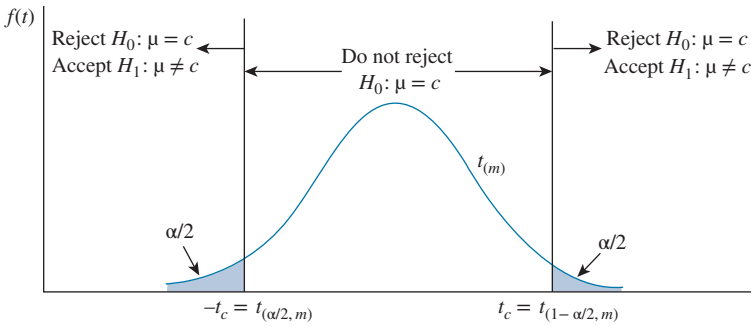


FIGURE C.7 **Rejection region for a test of $H_0: \mu = c$ against $H_1: \mu \neq c$.**

EXAMPLE C.11 | One-tail Test Using the Hip Data

Let us illustrate by testing the null hypothesis that the population hip size is 16.5 inches, against the alternative that it is *greater* than 16.5 inches. The following five-step format is recommended.

1. The null hypothesis is $H_0: \mu = 16.5$. The alternative hypothesis is $H_1: \mu > 16.5$.
2. The test statistic $t = \left(\overline{Y} - 16.5\right) \Big/ \left(\hat{\sigma}/\sqrt{N}\right) \sim t_{(N-1)}$ *if the null hypothesis is true*.
3. Let us select the level of significance $\alpha = 0.05$. The critical value $t_c = t_{(0.95,49)} = 1.6766$ for a t-distribution with $N - 1 = 49$ degrees of freedom. Thus we will reject the null hypothesis in favor of the alternative if $t \geq 1.68$.
4. Using the hip data, the estimate of μ is $\overline{y} = 17.1582$, with estimated variance $\hat{\sigma}^2 = 3.2653$, so $\hat{\sigma} = 1.807$. The value of the test statistic is
$$t = \frac{17.1582 - 16.5}{1.807/\sqrt{50}} = 2.5756$$
5. *Conclusion*: Since $t = 2.5756 > 1.68$, we *reject* the null hypothesis. The sample information we have is *incompatible* with the hypothesis that $\mu = 16.5$. We accept the alternative that the population mean hip size is greater than 16.5 inches, at the $\alpha = 0.05$ level of significance.

EXAMPLE C.12 | Two-tail Test Using the Hip Data

Let us test the null hypothesis that the population hip size is 17 inches, against the alternative that it is *not equal to* 17 inches. The steps of the test are

1. The null hypothesis is $H_0: \mu = 17$. The alternative hypothesis is $H_1: \mu \neq 17$.
2. The test statistic $t = \left(\overline{Y} - 17\right) \Big/ \left(\hat{\sigma}/\sqrt{N}\right) \sim t_{(N-1)}$ *if the null hypothesis is true*.
3. Let us select the level of significance $\alpha = 0.05$. In a two-tail test $\alpha/2 = 0.025$ of probability is allocated to each tail of the distribution. The critical value is the 97.5 percentile of the t-distribution, which leaves 2.5% of the probability in the upper tail, $t_c = t_{(0.975,49)} = 2.01$ for a t-distribution with $N - 1 = 49$ degrees of freedom. Thus, we will reject the null hypothesis in favor of the alternative if $t \geq 2.01$ or if $t \leq -2.01$.
4. Using the hip data, the estimate of μ is $\overline{y} = 17.1582$, with estimated variance $\hat{\sigma}^2 = 3.2653$, so $\hat{\sigma} = 1.807$. The value of the test statistic is
$$t = (17.1582 - 17) \Big/ \left(1.807/\sqrt{50}\right) = 0.6191.$$
5. *Conclusion*: Since $-2.01 < t = 0.6191 < 2.01$ we *do not reject* the null hypothesis. The sample information we have is *compatible* with the hypothesis that the population mean hip size $\mu = 17$.

Warning

Care must be taken when interpreting the outcome of a statistical test. One of the basic precepts of hypothesis testing is that finding a sample value of the test statistic in the nonrejection region does not make the null hypothesis true! Suppose another null hypothesis is $H_0: \mu = c^*$, where c^* is "close" to c. If we fail to reject the hypothesis $\mu = c$, then we will likely fail to reject the hypothesis that $\mu = c^*$. In the example above, at the $\alpha = 0.05$ level, we fail to reject the hypothesis that μ is 17, 16.8, 17.2, or 17.3. In fact, in any problem there are many hypotheses that we would fail to reject, but that does not make any of them true. The weaker statements "we do not reject the null hypothesis" or "we fail to reject the null hypothesis" do not send a misleading message.

C.6.5 The *p*-Value

When reporting the outcome of statistical hypothesis tests it has become common practice to report the ***p*-value** of the test. If we have the *p*-value of a test, p, we can determine the outcome of the test by comparing the *p*-value to the chosen level of significance, α, *without* looking up or calculating the critical values ourselves. The rule is

> ***p*-Value Rule**
>
> Reject the null hypothesis when the *p*-value is less than, or equal to, the level of significance α. That is, if $p \leq \alpha$ then reject H_0. If $p > \alpha$, then do not reject H_0.

If you have chosen the level of significance to be $\alpha = 0.01, 0.05, 0.10$, or any other value, you can compare it to the *p*-value of a test and then reject, or not reject, without checking the critical value t_c.

How the *p*-value is computed depends on the alternative. If t is the calculated value (not the critical value t_c) of the t-statistic with $N - 1$ degrees of freedom, then

- if $H_1 : \mu > c$, p = probability to the right of t
- if $H_1 : \mu < c$, p = probability to the left of t
- $H_1 : \mu \neq c$, p = *sum* of probabilities to the right of $|t|$ and to the left of $-|t|$

The direction of the alternative indicates the tail(s) of the distribution in which the *p*-value falls.

EXAMPLE C.13 | One-tail Test *p*-value: The Hip Data

In Example C.11 we used the hip data to test $H_0 : \mu = 16.5$ against $H_1 : \mu > 16.5$. The calculated t-statistic value was $t = 2.5756$. In this case, since the alternative is "greater than" (>), the *p*-value of this test is the probability that a t-random variable with $N - 1 = 49$ degrees of freedom is greater than 2.5756. This probability value cannot be found in the usual t-table of critical values, but it is easily found using the computer. Statistical software packages, and spreadsheets such as Excel, have simple commands to evaluate the *cumulative distribution function* (*cdf*) (see the Probability Primer, Section P.2) for a variety of probability distributions. If $F_X(x)$ is the *cdf* for a random variable X, then for any value $x = c$, $P[X \leq c] = F_X(c)$. Given such a function for the t-distribution, we compute the desired *p*-value as

$$p = P\left(t_{(49)} \geq 2.576\right) = 1 - P\left(t_{(49)} \leq 2.576\right) = 0.0065$$

Given the *p*-value, we can immediately conclude that at $\alpha = 0.01$ or 0.05 we reject the null hypothesis in favor of the alternative, but if $\alpha = 0.001$ we would not reject the null hypothesis.

The logic of the *p*-value rule is shown in Figure C.8. If 0.0065 of the probability lies to the right of $t = 2.5756$, then the critical value t_c that leaves a probability of $\alpha = 0.01\left(t_{(0.99,49)}\right)$ or $\alpha = 0.05\left(t_{(0.95,49)}\right)$ in the tail must be to the left of 2.5756. In this case, when the *p*-value $\leq \alpha$, it must be true that $t \geq t_c$, and we should reject the null hypothesis for either of these levels of significance. On the other hand, it must be true that the critical value for $\alpha = 0.001$ must fall to the right of 2.5756, meaning that we should not reject the null hypothesis at this level of significance.

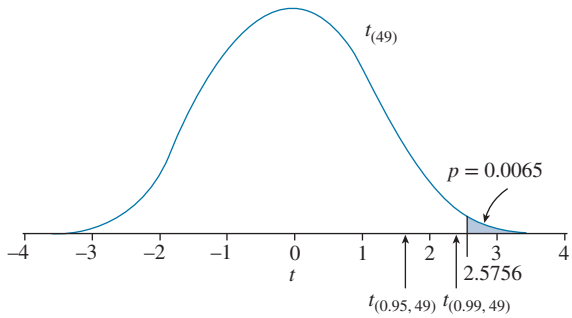


FIGURE C.8 ***p*-value for a right-tail test.**

EXAMPLE C.14 | Two-Tail Test *p*-Value: The Hip Data

For a two-tail test, the rejection region is in the two tails of the *t*-distribution, and the *p*-value must similarly be calculated in the two tails of the distribution. For the hip data, we tested the null hypothesis $H_0: \mu = 17$ against $H_1: \mu \neq 17$, yielding the test statistic value $t = 0.6191$. The *p*-value is

$$p = P\left[t_{(49)} \geq 0.6191\right] + P\left[t_{(49)} \leq -0.6191\right]$$
$$= 2 \times 0.2694 = 0.5387$$

Since the *p*-value $= 0.5387 > \alpha = 0.05$, we do not reject the null hypothesis $H_0: \mu = 17$ at $\alpha = 0.05$ or any other common level of significance. The two-tail *p*-value is shown in Figure C.9.

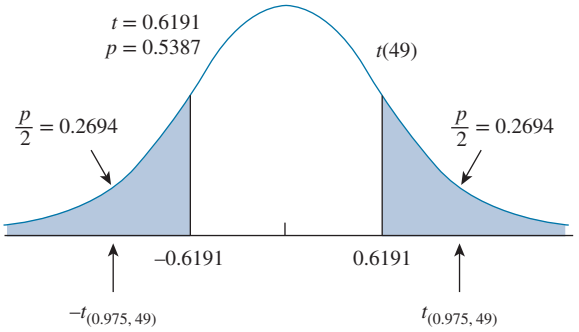


FIGURE C.9 **The *p*-value for a two-tail test.**

C.6.6 A Comment on Stating Null and Alternative Hypotheses

A statistical test procedure cannot prove the truth of a null hypothesis. When we fail to reject a null hypothesis, all the hypothesis test can establish is that the information in a sample of data is *compatible* with the null hypothesis. On the other hand, a statistical test can lead us to *reject* the null hypothesis, with only a small probability, α, of rejecting the null hypothesis when it is actually true. Thus rejecting a null hypothesis is a stronger conclusion than failing to reject it.

The null hypothesis is usually stated in such a way that if our theory is correct, then we will reject the null hypothesis. For example, our airplane seat designer has been operating under the assumption (the maintained or null hypothesis) that the population mean hip width is 16.5 inches. Casual observation suggests that people are getting larger all the time. If we are larger, and if the airline wants to continue to accommodate the same percentage of the population, then the seat widths must be increased. This costly change should be undertaken only if there is statistical evidence that the population hip size is indeed larger. When using a hypothesis test we would like to find out whether there is statistical evidence against our current "theory," or whether the data are compatible with it. With this goal, we set up the null hypothesis that the population mean is 16.5 inches, $H_0: \mu = 16.5$, against the alternative that it is greater than 16.5 inches, $H_1: \mu > 16.5$. In this case, if we reject the null hypothesis, we have shown that there has been a "statistically significant" increase in hip width.

You may view the null hypothesis to be too limited in this case, since it is feasible that the population mean hip width is now smaller than 16.5 inches. The hypothesis test of the null hypothesis $H_0: \mu \leq 16.5$ against the alternative hypothesis $H_1: \mu > 16.5$ is exactly the same as the test for $H_0: \mu = 16.5$ against the alternative hypothesis $H_1: \mu > 16.5$. The test statistic and rejection region are exactly the same. For a one-tail test you can form the null hypothesis in either of these ways.

Finally, it is important to set up the null and alternative hypotheses before you analyze or even collect the sample of data. Failing to do so can lead to errors in formulating the alternative hypothesis. Suppose that we wish to test whether $\mu > 16.5$ and the sample mean is $\bar{y} = 15.5$. Does that mean we should set up the alternative $\mu < 16.5$, to be consistent with the estimate? The answer is no. The alternative is formed to state the conjecture that we wish to establish, $\mu > 16.5$.

C.6.7 Type I and Type II Errors

Whenever we reject—or do not reject—a null hypothesis, there is a chance that we may be making a mistake. This is unavoidable. In any hypothesis testing situation, there are two ways that we can make a correct decision and two ways that we can make an incorrect decision.

Correct Decisions

The null hypothesis is *false* and we decide to *reject* it.

The null hypothesis is *true* and we decide *not* to reject it.

Incorrect Decisions

The null hypothesis is *true* and we decide to *reject* it (a Type I error).

The null hypothesis is *false* and we decide *not* to reject it (a Type II error).

When we reject the null hypothesis we risk what is called a **Type I error**. The probability of a Type I error is α, the level of significance of the test. When the null hypothesis is true, the t-statistic falls in the rejection region with probability α. Thus hypothesis tests will *reject* a true hypothesis $100\alpha\%$ of the time. The good news here is that we can control the probability of a Type I error by choosing the level of significance of the test, α.

We risk a **Type II error** when we do not reject the null hypothesis. Hypothesis tests will lead us to fail to reject null hypotheses that are false with a certain probability. The magnitude of the probability of a Type II error is not under our control and cannot be computed, because it depends on the true value of μ, which is unknown. However, we do know that

- The probability of a Type II error varies inversely with the level of significance of the test, α, which is the probability of a Type I error. If you choose to make α smaller, the probability of a Type II error increases.
- If the null hypothesis is $\mu = c$, and if the true (unknown) value of μ is *close* to c, then the probability of a Type II error is high.
- The larger the sample size N, the lower the probability of a Type II error, given a level of Type I error α.

An easy to remember example of the difference between Type I and Type II errors is from the U.S. legal system. In a trial, a person is presumed innocent. This is the "null" hypothesis, the alternative hypothesis being that the person is guilty. If we convict an innocent person, then we have rejected a null hypothesis that is true, committing a Type I error. If we fail to convict a guilty person, failing to reject the false null hypothesis, then we commit a Type II error. Which is the more costly error in this context? Is it better to send an innocent person to jail, or to let a guilty person go free? It is better in this case to make the probability of a Type I error very small.

C.6.8 A Relationship Between Hypothesis Testing and Confidence Intervals

There is an algebraic relationship between two-tail hypothesis tests and confidence interval estimates that is sometimes useful. Suppose that we are testing the null hypothesis $H_0 : \mu = c$ against the alternative $H_1 : \mu \neq c$. If we fail to reject the null hypothesis at the α level of significance,

then the value c will fall within a $100(1-\alpha)\%$ confidence interval estimate of μ. Conversely, if we reject the null hypothesis, then c will fall outside the $100(1-\alpha)\%$ confidence interval estimate of μ. This algebraic relationship is true because we fail to reject the null hypothesis when $-t_c \le t \le t_c$, or when

$$-t_c \le \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \le t_c$$

which when rearranged becomes

$$\overline{Y} - t_c \frac{\hat{\sigma}}{\sqrt{N}} \le c \le \overline{Y} + t_c \frac{\hat{\sigma}}{\sqrt{N}}$$

The endpoints of this interval are the same as the endpoints of a $100(1-\alpha)\%$ confidence interval estimate of μ. Thus for any value of c within the confidence interval, we do not reject $H_0: \mu = c$ against the alternative $H_1: \mu \ne c$. For any value of c outside the confidence interval, we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu \ne c$.

This relationship can be handy if you are given only a confidence interval and want to determine what the outcome of a two-tail test would be.

C.7 Some Other Useful Tests

In this section we very briefly summarize some additional tests. These tests are not only useful in and of themselves, but also illustrate the use of test statistics with chi-square and F-distributions. These distributions were introduced in Appendix B.3.

C.7.1 Testing the Population Variance

Let Y be a normally distributed random variable, $Y \sim N(\mu, \sigma^2)$. Assume that we have a random sample of size N from this population, $Y_1, Y_2, \ldots, Y_N$. The estimator of the population mean is $\overline{Y} = \sum Y_i/N$, and the unbiased estimator of the population variance is $\hat{\sigma}^2 = \sum\left(Y_i - \overline{Y}\right)^2/(N-1)$. To test the null hypothesis $H_0: \sigma^2 = \sigma_0^2$, we use the test statistic

$$V = \frac{(N-1)\,\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{(N-1)}$$

If the null hypothesis is true, then the test statistic has the indicated chi-square distribution with $N-1$ degrees of freedom. If the alternative hypothesis is $H_1: \sigma^2 > \sigma_0^2$, then we carry out a one-tail test. If we choose the level of significance $\alpha = 0.05$, then the null hypothesis is rejected if $V \ge \chi^2_{(0.95,N-1)}$, where $\chi^2_{(0.95,N-1)}$ is the 95th percentile of the chi-square distribution with $N-1$ degrees of freedom. These values can be found in Statistical Table 3, or computed using statistical software. If the alternative hypothesis is $H_1: \sigma^2 \ne \sigma_0^2$, then we carry out a two-tail test, and the null hypothesis is rejected if $V \ge \chi^2_{(0.975,N-1)}$ or if $V \le \chi^2_{(0.025,N-1)}$. The chi-square distribution is skewed, with a long tail to the right, so we cannot use the properties of symmetry when determining the left- and right-tail critical values.

C.7.2 Testing the Equality of Two Population Means

Let two normal populations be denoted by $N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$. In order to estimate and test the difference between means, $\mu_1 - \mu_2$, we must have random samples of data from each of the two populations. We draw a sample of size N_1 from the first population, and a sample of size N_2 from the second population. Using the first sample we obtain the sample mean $\overline{Y}_1$ and sample

variance $\hat{\sigma}_1^2$; from the second sample we obtain $\overline{Y}_2$ and $\hat{\sigma}_2^2$. How the null hypothesis $H_0: \mu_1 - \mu_2 = c$ is tested depends on whether the two population variances are equal or not.

Case 1: *Population variances are equal* If the population variances are equal, so that $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$, then we use information in both samples to estimate the common value σ_p^2. This "pooled variance estimator" is

$$\hat{\sigma}_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}$$

If the null hypothesis $H_0: \mu_1 - \mu_2 = c$ is true, then

$$t = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - c}{\sqrt{\hat{\sigma}_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim t_{(N_1+N_2-2)}$$

As usual, we can construct a one-sided alternative, such as $H_1: \mu_1 - \mu_2 > c$, or the two-sided alternative $H_1: \mu_1 - \mu_2 \neq c$.

Case 2: *Population variances are unequal* If the population variances are not equal, then we cannot use the pooled variance estimate. Instead, we use

$$t^* = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - c}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

The exact distribution of this test statistic is neither normal nor the usual t-distribution. The distribution of t^* can be approximated by a t-distribution with degrees of freedom

$$df = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\frac{\left(\hat{\sigma}_1^2/N_1\right)^2}{N_1 - 1} + \frac{\left(\hat{\sigma}_2^2/N_2\right)^2}{N_2 - 1}\right)}$$

This is one of several approximations that appear in the statistics literature, and your software may well use a different one.

C.7.3 Testing the Ratio of Two Population Variances

Given two normal populations, denoted by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, we can test the null hypothesis $H_0: \sigma_1^2/\sigma_2^2 = 1$. If the null hypothesis is true, then the population variances are equal. The test statistic is derived from the results that $(N_1 - 1)\hat{\sigma}_1^2/\sigma_1^2 \sim \chi^2_{(N_1-1)}$ and $(N_2 - 1)\hat{\sigma}_2^2/\sigma_2^2 \sim \chi^2_{(N_2-1)}$. In Appendix B.3.8 we define an F-random variable, which is formed by taking the ratio of two independent chi-square random variables that have been divided by their degrees of freedom. In this case, the relevant ratio is

$$F = \frac{\frac{(N_1 - 1)\,\hat{\sigma}_1^2/\sigma_1^2}{(N_1 - 1)}}{\frac{(N_2 - 1)\,\hat{\sigma}_2^2/\sigma_2^2}{(N_2 - 1)}} = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2} \sim F_{(N_1-1, N_2-1)}$$

If the null hypothesis $H_0: \sigma_1^2/\sigma_2^2 = 1$ is true then the test statistic is $F = \hat{\sigma}_1^2/\hat{\sigma}_2^2$, which has an F-distribution with $N_1 - 1$ numerator and $N_2 - 1$ denominator degrees of freedom. If the

alternative hypothesis is $H_1: \sigma_1^2/\sigma_2^2 \neq 1$, then we carry out a two-tail test. If we choose level of significance $\alpha = 0.05$, then we reject the null hypothesis if $F \geq F_{(0.975, N_1-1, N_2-1)}$ or if $F \leq F_{(0.025, N_1-1, N_2-1)}$, where $F_{(\alpha, N_1-1, N_2-1)}$ denotes the 100α-percentile of the F-distribution with the specified degrees of freedom. If the alternative is one sided, $H_1: \sigma_1^2/\sigma_2^2 > 1$, then we reject the null hypothesis if $F \geq F_{(0.95, N_1-1, N_2-1)}$.

C.7.4 Testing the Normality of a Population

The tests for means and variances we have developed began with the assumption that the populations were normally distributed. Two questions immediately arise. How well do the tests work when the population is not normal? Can we test for the normality of a population? The answer to the first question is that the tests work pretty well even if the population is not normal, so long as samples are sufficiently large. How large must the samples be? There is no easy answer, since it depends on how "nonnormal" the populations are. The answer to the second question is yes, we can test for normality. Statisticians have been vitally interested in this question for a long time, and a variety of tests have been developed, but the tests and underlying theory are very complicated and far outside the scope of this book.

However, we can present a test that is slightly less ambitious. The normal distribution is symmetric and has a bell shape with a peakedness and tail thickness leading to a kurtosis of three. Thus we can test for departures from normality by checking the skewness and kurtosis from a sample of data. If skewness is not close to zero, or if kurtosis is not close to three, then we reject the normality of the population. In Section C.4.2 we developed sample measures of skewness and kurtosis as

$$\widehat{skewness} = S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{kurtosis} = K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

The **Jarque–Bera** test statistic allows a joint test of these two characteristics,

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

If the true distribution is symmetric and has kurtosis three, which includes the normal distribution, then the JB test statistic has a chi-square distribution with two degrees of freedom if the sample size is sufficiently large. If $\alpha = 0.05$, then the critical value of the $\chi^2_{(2)}$ distribution is 5.99. We reject the null hypothesis and conclude that the data are nonnormal if $JB \geq 5.99$. If we reject the null hypothesis, then we know that the data have nonnormal characteristics, but we do not know what distribution the population might have.

EXAMPLE C.15 | Testing the Normality of the Hip Data

For the hip data, skewness and kurtosis measures were estimated in Example C.6. Plugging these values into the JB test statistic formula we obtain

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$
$$= \frac{50}{6}\left((-0.0138)^2 + \frac{(2.3315-3)^2}{4}\right) = 0.9325$$

Since $JB = 0.9325$ is less than the critical value 5.99, we conclude that we cannot reject the normality of the hip data. The p-value for this test is the tail area of a $\chi^2_{(2)}$ -distribution to the right of 0.9325,

$$p = P\left[\chi^2_{(2)} \geq 0.9325\right] = 0.6273$$

C.8 Introduction to Maximum Likelihood Estimation[1]

Maximum likelihood estimation is a powerful procedure that can be used when the population distribution is known. In this section we introduce the concept with a very simple but revealing example.

EXAMPLE C.16 | The "Wheel of Fortune" Game: $p = 1/4$ or $3/4$

Consider the following "Wheel of Fortune" game. You are a contestant faced with two wheels, each of which is partly shaded and partly nonshaded (see Figure C.10). Suppose that after spinning a wheel, you win if a pointer is in the shaded area, and you lose if the pointer is in the nonshaded area. On wheel A 25% of the area is shaded so that the probability of winning is 1/4. On wheel B 75% of the area is shaded so that the probability of winning is 3/4. The game that you must play is this. One of the wheels is chosen and spun three times, with outcomes WIN, WIN, LOSS. You *do not* know which wheel was chosen, and must pick which wheel was spun. Which would you select?

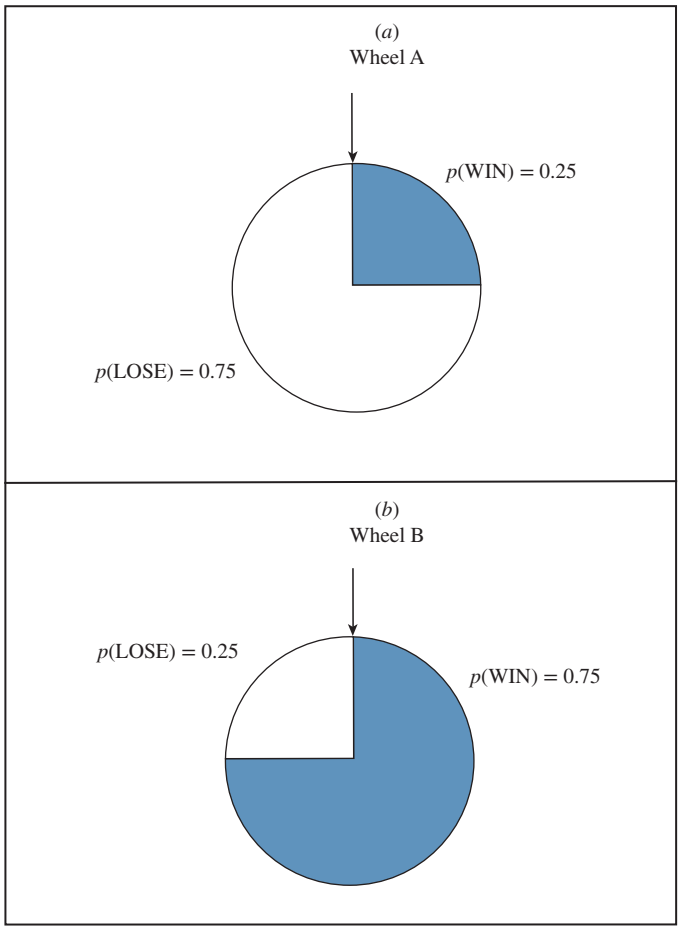


FIGURE C.10 **Wheel of fortune game.**

[1]This section contains some advanced material.

One intuitive approach is the following: let p denote the probability of winning on one spin of a wheel. Choosing between wheels A and B means choosing between $p = 1/4$ and $p = 3/4$. You are estimating p, but there are only two possible estimates, and you must choose based on the observed data. Let us compute the probability of each sequence of outcomes for each of the wheels.

For wheel A, with $p = 1/4$, the probability of observing WIN, WIN, LOSS is

$$\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64} = 0.0469$$

That is, the probability, or **likelihood**, of observing the sequence WIN, WIN, LOSS when $p = 1/4$ is 0.0469.

For wheel B, with $p = 3/4$, the probability of observing WIN, WIN, LOSS is

$$\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{9}{64} = 0.1406$$

The probability, or likelihood, of observing the sequence WIN, WIN, LOSS when $p = 3/4$ is 0.1406.

If we had to choose wheel A or B based on the available data, we would choose wheel B because it has a higher probability of having produced the observed data. It is more *likely* that wheel B was spun than wheel A, and $\hat{p} = 3/4$ is called the **maximum likelihood estimate** of p. The **maximum likelihood principle** seeks the parameter values that maximize the probability, or likelihood, of observing the outcomes actually obtained.

EXAMPLE C.17 | The "Wheel of Fortune" Game: $0 < p < 1$

Now suppose p can be any probability between zero and one, not just 1/4 or 3/4. We have one wheel with a proportion of it shaded, which is the probability of WIN, but we do not know the proportion. In three spins we observe WIN, WIN, LOSS. What is the most likely value of p? The probability of observing WIN, WIN, LOSS is the likelihood L and is

$$L(p) = p \times p \times (1 - p) = p^2 - p^3 \qquad \text{(C.17)}$$

The likelihood L depends on the unknown probability p of a WIN, which is why we have given it the notation $L(p)$, indicating a functional relationship. We would like to find the value of p that maximizes the likelihood of observing the outcomes actually obtained. The graph of the likelihood function (C.17) and the choice of p that maximizes this function is shown in Figure C.11. The maximizing value is denoted as $\hat{p}$ and is called the maximum likelihood estimate of p. To find this value of p we can use calculus. Differentiate $L(p)$ with respect to p,

$$\frac{dL(p)}{dp} = 2p - 3p^2$$

Set this derivative to zero:

$$2p - 3p^2 = 0 \Rightarrow p(2 - 3p) = 0$$

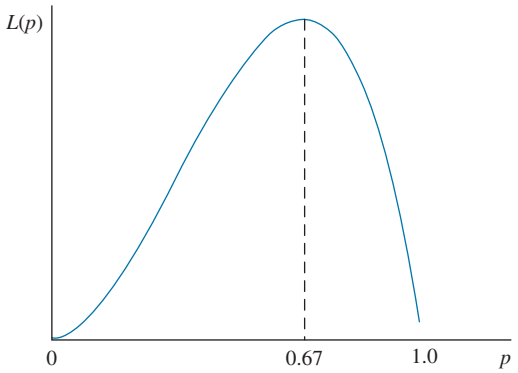


FIGURE C.11 **A likelihood function.**

There are two solutions to this equation, $p = 0$ or $p = 2/3$. The value that maximizes $L(p)$ is $\hat{p} = 2/3$, which is the maximum likelihood estimate. That is, of all possible values of p, between zero and one, the value that maximizes the probability of observing two wins and one loss (the order does not matter) is $\hat{p} = 2/3$.

Can we derive a more general formula that can be used for any observed data? In Appendix B.3.1 we introduced the Bernoulli distribution. Let us define the random variable X that takes the values $x = 1$ (WIN) and $x = 0$ (LOSS) with probabilities p and $1 - p$. The probability function for this random variable can be written in mathematical form as

$$P(X = x) = f(x|p) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

If we spin the "wheel" N times we observe N sample values $x_1, x_2, \ldots, x_N$. Assuming that the spins are independent, we can form the joint probability function

$$\begin{aligned} f(x_1, \ldots, x_N|p) &= f(x_1|p) \times \cdots \times f(x_N|p) \\ &= p^{\Sigma x_i}(1-p)^{N-\Sigma x_i} \\ &= L(p|x_1, \ldots, x_N) \end{aligned} \tag{C.18}$$

The joint probability function gives the probability of observing a specific set of outcomes, and it is a generalization of (C.17). In the last line we have indicated that the joint probability function is algebraically equivalent to the **likelihood function** $L(p|x_1, \ldots, x_N)$. The notation emphasizes that the likelihood function depends upon the unknown probability p *given* the sample outcomes, which we observe. For notational simplicity we will continue to denote the likelihood function as $L(p)$.

EXAMPLE C.18 | The "Wheel of Fortune" Game: Maximizing the Log-likelihood

In the "Wheel of Fortune" game, the maximum likelihood estimate is that value of p that maximizes $L(p)$. To find this estimate using calculus we use a trick to simplify the algebra. The value of p that maximizes $L(p) = p^2(1-p)$ is the same value of p that maximizes the **log-likelihood function** $\ln L(p) = 2\ln(p) + \ln(1-p)$, where "ln" is the natural logarithm. The plot of the log-likelihood function is shown in Figure C.12. Compare Figures C.11 and C.12. The maximum of the likelihood function is $L(\hat{p}) = 0.1481$. The maximum of the log-likelihood function is $\ln L(\hat{p}) = -1.9095$. Both of these maximum values occur at $\hat{p} = 2/3 = 0.6667$.

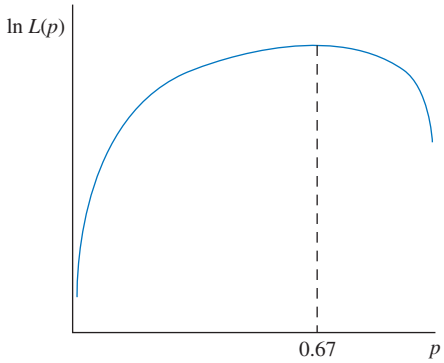


FIGURE C.12 **A log-likelihood function.**

The trick in Example C.18 works for all likelihood and log-likelihood functions and their parameters, so when you see maximum likelihood estimation being discussed it will always be in terms of maximizing the log-likelihood function. For the general problem we are considering, the log-likelihood function is the logarithm of (C.18)

$$\begin{aligned} \ln L(p) &= \sum_{i=1}^{N} \ln\left[f(x_i|p)\right] \\ &= \left(\sum_{i=1}^{N} x_i\right)\ln(p) + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1-p) \end{aligned} \tag{C.19}$$

The first derivative is

$$\frac{d\ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{N - \sum x_i}{1-p}$$

Setting this to zero and replacing p by $\hat{p}$ to denote the value that maximizes $\ln L(p)$ yields

$$\frac{\sum x_i}{\hat{p}} - \frac{N - \sum x_i}{1-\hat{p}} = 0$$

To solve this equation, multiply both sides by $\hat{p}(1-\hat{p})$. This gives

$$(1-\hat{p})\sum x_i - \hat{p}\left(N - \sum x_i\right) = 0$$

Finally, solving for $\hat{p}$ yields

$$\hat{p} = \frac{\sum x_i}{N} = \bar{x} \tag{C.20}$$

The estimator $\hat{p}$ is the **sample proportion**; $\sum x_i$ is the total number of 1s (wins) out of N spins. As you can see, $\hat{p}$ is also the sample mean of x_i. This result is completely general. Any time we have two outcomes that can occur with probabilities p and $1 - p$, then the maximum likelihood estimate based on a sample of N observations is the sample proportion (C.20).

EXAMPLE C.19 | Estimating a Population Proportion

This estimation strategy can be used if you are a pollster trying to estimate the proportion of the population intending to vote for candidate A rather than candidate B, a medical researcher who wishes to estimate the proportion of the population having a particular defective gene, or a marketing researcher trying to discover whether the population of customers prefers a blue box or a green box for their morning cereal. Suppose in this latter case that you select 200 cereal consumers at random and ask whether they prefer blue boxes or green. If 75 prefer a blue box, then we would estimate that the population proportion preferring blue is $\hat{p} = \sum x_i/N = 75/200 = 0.375$. Thus, we estimate that 37.5% of the population prefers a blue box.

C.8.1 Inference with Maximum Likelihood Estimators

If we use maximum likelihood estimation, how do we perform hypothesis tests and construct confidence intervals? The answers to these questions are found in some remarkable properties of estimators obtained using maximum likelihood methods. Let us consider a general problem. Let X be a random variable (either discrete or continuous) with a probability density function $f(x|\theta)$, where θ is an unknown parameter. The log-likelihood function, based on a random sample $x_1, \ldots, x_N$ of size N, is

$$\ln L(\theta) = \sum_{i=1}^{N} \ln\left[f\left(x_i|\theta\right)\right]$$

If the probability density function of the random variable involved is relatively smooth, and if certain other technical conditions hold, then in large samples the maximum likelihood estimator $\hat{\theta}$ of a parameter θ has a probability distribution that is approximately normal, with expected value θ and a variance $V = \text{var}\left(\hat{\theta}\right)$ that we will discuss in a moment. That is, we can say

$$\hat{\theta} \overset{a}{\sim} N(\theta, V) \tag{C.21}$$

where the symbol $\overset{a}{\sim}$ denotes "asymptotically distributed." The word "asymptotic" refers to estimator properties when the sample size N becomes large, or as $N \to \infty$. To say that an estimator is asymptotically normal means that its probability distribution, which may be unknown when samples are small, becomes approximately normal in large samples. This is analogous to the central limit theorem we discussed in Section C.3.4.

Based on the normality result in (C.21) it will not surprise you that we can immediately construct a t-statistic and obtain both a confidence interval and a test statistic from it. Specifically, if we wish to test the null hypothesis $H_0: \theta = c$ against a one-tail or two-tail alternative hypothesis, then we can use the test statistic

$$t = \frac{\hat{\theta} - c}{\text{se}\left(\hat{\theta}\right)} \overset{a}{\sim} t_{(N-1)} \tag{C.22}$$

If the null hypothesis is true, then this t-statistic has a distribution that can be approximated by a t-distribution with $N - 1$ degrees of freedom in large samples. The mechanics of carrying out the hypothesis test are exactly those in Section C.6.

If t_c denotes the 100(1 – α/2)-percentile $t_{(1-\alpha/2,N-1)}$, then a 100(1 − α)% confidence interval for θ is

$$\hat{\theta} \pm t_c \text{se}\left(\hat{\theta}\right)$$

This confidence interval is interpreted just like those in Section C.5.

Remark

These asymptotic results in (C.21) and (C.22) hold only in large samples. We have indicated that the distribution of the test statistic can be approximated by a t-distribution with $N - 1$ degrees of freedom. If N is truly large, then the $t_{(N-1)}$-distribution converges to the standard normal distribution $N(0, 1)$ and the 100(1 − α/2)-percentile value $t_{(1-\alpha/2,N-1)}$ converges to the corresponding percentile from the standard normal distribution. Asymptotic results are used, rightly or wrongly, when the sample size N may not be large. We prefer using the t-distribution critical values, which are adjusted for small samples by the degrees of freedom correction, when obtaining interval estimates and carrying out hypothesis tests.

C.8.2 The Variance of the Maximum Likelihood Estimator

A key ingredient in both the test statistic and confidence interval expressions is the standard error $\text{se}\left(\hat{\theta}\right)$. Where does this come from? Standard errors are square roots of estimated variances. The part we have delayed discussing until now is how we find the variance of the maximum likelihood estimator, $V = \text{var}\left(\hat{\theta}\right)$. The variance V is given by the inverse of the negative expectation of the second derivative of the log-likelihood function,

$$V = \text{var}\left(\hat{\theta}\right) = \left[-E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)\right]^{-1} \tag{C.23}$$

This looks quite intimidating, and you can see why we put it off. What does this mean? First of all, the second derivative measures the curvature of the log-likelihood function. A second derivative is literally the derivative of the derivative see Appendix A.3.3. A single derivative, the first, measures the slope of a function or the rate of change of the function. The second derivative measures the rate of change of the slope. To obtain a maximum of the log-likelihood function, it must be an "inverted bowl" shape, like those shown in Figure C.13.

At any point to the left of the maximum point, the slope of the log-likelihood function is positive. At any point to the right of the maximum, the slope is negative. As we progress from left to right the slope is *decreasing* (becoming less positive or more negative), so that the second derivative must be negative. A larger absolute magnitude of the second derivative implies a

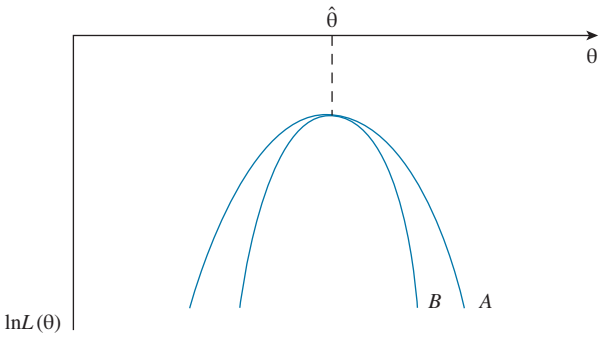


FIGURE C.13 **The log-likelihood functions.**

more rapidly changing slope, indicating a more sharply curved log-likelihood. This is important. In Figure C.13 the two log-likelihood functions A and B have the same maximizing value $\hat{\theta}$. Imagine yourself a climber who is trekking up one of these mountains. For which mountain is the summit most clearly defined? For log-likelihood B, the summit is a sharp peak, and its maximum is more easily located than that for log-likelihood A. The sharper peak has less "wiggle room" at the summit. The smaller amount of wiggle room means that there is less uncertainty as to the location of the maximizing value $\hat{\theta}$; in estimation terminology, less uncertainty means greater precision, and a smaller variance. The more sharply curved log-likelihood function, the one whose second derivative is larger in absolute magnitude, leads to more precise maximum likelihood estimation, and to a maximum likelihood estimator with smaller variance. Thus the variance V of the maximum likelihood estimator is inversely related to the (negative) second derivative. The expected value "E" must be present because this quantity depends on the data and is thus random, so we average over all possible data outcomes.

C.8.3 The Distribution of the Sample Proportion

It is time for an example. At the beginning of Section C.8 we introduced a random variable X that takes the values $x = 1$ and $x = 0$ with probabilities p and $1 - p$. It has log-likelihood given in (C.19). In this problem the parameter θ that we are estimating is the population proportion p, the proportion of $x = 1$ values in the population. We already know that the maximum likelihood estimator of p is the sample proportion $\hat{p} = \sum x_i / N$. The second derivative of the log-likelihood function (C.19) is

$$\frac{d^2 \ln L(p)}{dp^2} = -\frac{\sum x_i}{p^2} - \frac{N - \sum x_i}{(1-p)^2} \tag{C.24}$$

To calculate the variance of the maximum likelihood estimator we need the "expected value" of expression (C.24). In the expectation we treat the x_i values as random because these values vary from sample to sample. The expected value of this discrete random variable is obtained using (P.9) in the probability primer:

$$E(x_i) = 1 \times P(x_i = 1) + 0 \times P(x_i = 0) = 1 \times p + 0 \times (1-p) = p$$

Then, using a generalization of (P.16) (the expected value of a sum is the sum of the expected values and constants can be factored out of expectations) we find the expected value of the second derivative as

$$\begin{aligned} E\left(\frac{d^2 \ln L(p)}{dp^2}\right) &= -\frac{\sum E(x_i)}{p^2} - \frac{N - \sum E(x_i)}{(1-p)^2} \\ &= -\frac{Np}{p^2} - \frac{N - Np}{(1-p)^2} \\ &= -\frac{N}{p(1-p)} \end{aligned}$$

The variance of the sample proportion, which is the maximum likelihood estimator of p, is then

$$V = \text{var}(\hat{p}) = \left[-E\left(\frac{d^2 \ln L(p)}{dp^2}\right)\right]^{-1} = \frac{p(1-p)}{N}$$

The **asymptotic distribution** of the sample proportion, which is valid in large samples, is

$$\hat{p} \overset{a}{\sim} N\left(p, \frac{p(1-p)}{N}\right)$$

To estimate the variance V we must replace the true population proportion by its estimate,

$$\hat{V} = \frac{\hat{p}(1-\hat{p})}{N}$$

The standard error that we need for hypothesis testing and confidence interval estimation is the square root of this estimated variance:

$$\text{se}(\hat{p}) = \sqrt{\hat{V}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

EXAMPLE C.20 | Testing a Population Proportion

As a numerical example, suppose a cereal company CEO conjectures that 40% of the population prefers a blue box. To test this hypothesis, we construct the null hypothesis $H_0: p = 0.4$ and use the two-tail alternative $H_1: p \neq 0.4$. If the null hypothesis is true, then the test statistic $t = (\hat{p} - 0.4)/\text{se}(\hat{p}) \overset{a}{\sim} t_{(N-1)}$. For a sample of size $N = 200$ the critical value from the t-distribution is $t_c = t_{(0.975,199)} = 1.97$. Therefore we reject the null hypothesis if the calculated value of $t \geq 1.97$ or $t \leq -1.97$. If 75 of the respondents prefer a blue box, then the sample proportion is $\hat{p} = 75/200 = 0.375$. The standard error of this estimate is

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \sqrt{\frac{0.375 \times 0.625}{200}} = 0.0342$$

The value of the test statistic is

$$t = \frac{\hat{p} - 0.4}{\text{se}(\hat{p})} = \frac{0.375 - 0.4}{0.0342} = -0.7303$$

This value is in the nonrejection region, $-1.97 < t = -0.7303 < 1.97$, so we do not reject the null hypothesis that $p = 0.4$. The sample data are compatible with the conjecture that 40% of the population prefer a blue box.

The 95% interval estimate of the population proportion p who prefer a blue box is

$$\hat{p} \pm 1.97\text{se}(\hat{p}) = 0.375 \pm 1.97(0.0342) = [0.3075, 0.4424]$$

We estimate that between 30.8% and 44.3% of the population prefer a blue box.

C.8.4 Asymptotic Test Procedures

When using maximum likelihood estimation, there are three test procedures that can be used, with the choice depending on which one is most convenient in a given case. The tests are *asymptotically equivalent* and will give the same result in large samples. Suppose that we are testing the null hypothesis $H_0: \theta = c$ against the alternative hypothesis $H_1: \theta \neq c$. In (C.22) we have the t-statistic for carrying out the test. How does this test really work? Basically it is measuring the distance $\hat{\theta} - c$ between the estimate of θ and the hypothesized value c. This distance is normalized by the standard error of $\hat{\theta}$ to adjust for how precisely we have estimated θ. If the distance between the estimate $\hat{\theta}$ and the hypothesized value c is large, then that is taken as evidence against the null hypothesis, and if the distance is large enough, we conclude that the null hypothesis is not true.

There are other ways to measure the distance between $\hat{\theta}$ and c that can be used to construct test statistics. Each of the three testing principles takes a different approach to measuring the distance between $\hat{\theta}$ and the hypothesized value.

The Likelihood Ratio (LR) Test Consider Figure C.14. A log-likelihood function is shown, along with the maximum likelihood estimate $\hat{\theta}$ and the hypothesized value c. Note that the distance between $\hat{\theta}$ and c is also reflected by the distance between the log-likelihood function value evaluated at the maximum likelihood estimate $\ln L\left(\hat{\theta}\right)$ and the log-likelihood function value evaluated at the hypothesized value $\ln L(c)$. We have labeled the difference between these two log-likelihood values (1/2)LR for a reason that will become clear. If the estimate $\hat{\theta}$ is close to c, then the difference between the log-likelihood values will be small. If $\hat{\theta}$ is far from c, then

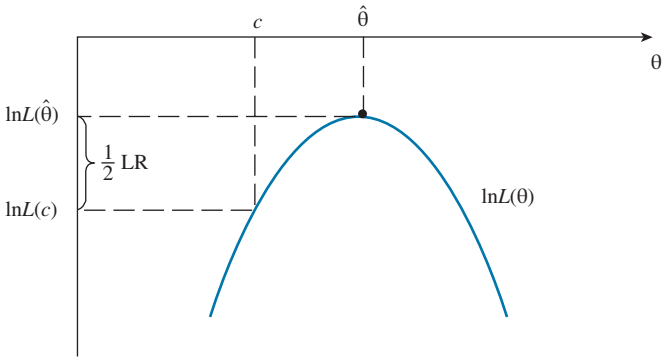


FIGURE C.14 **The likelihood ratio test.**

the difference between the log-likelihood values will be large. This observation leads us to the **likelihood ratio statistic**, which is twice the difference between $\ln L\left(\hat{\theta}\right)$ and $\ln L(c)$,

$$\text{LR} = 2\left[\ln L\left(\hat{\theta}\right) - \ln L(c)\right] \tag{C.25}$$

Based on some advanced statistical theory, it can be shown that if the null hypothesis is true, then the LR test statistic has a chi-square distribution (see Appendix B.3.6) with $J = 1$ degree of freedom. In more general contexts J is the number of hypotheses being tested and it can be greater than 1. If the null hypothesis is not true, then the LR test statistic becomes large. We reject the null hypothesis at the α level of significance if $\text{LR} \geq \chi^2_{(1-\alpha,J)}$, where $\chi^2_{(1-\alpha,J)}$ is the $100(1-\alpha)$-percentile of a chi-square distribution with J degrees of freedom, as shown in Figure C.15. The 90th, 95th, and 99th percentile values of the chi-square distribution for various degrees of freedom are given in Statistical Table 3.

When estimating a population proportion p the log-likelihood function is given by (C.19). The value of p that maximizes this function is $\hat{p} = \sum x_i/N$. Thus, the maximum value of the log-likelihood function is

$$\begin{aligned}\ln L(\hat{p}) &= \left(\sum_{i=1}^{N} x_i\right)\ln\hat{p} + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1-\hat{p}) \\ &= N\hat{p}\ln\hat{p} + (N - N\hat{p})\ln(1-\hat{p}) \\ &= N\left[\hat{p}\ln\hat{p} + (1-\hat{p})\ln(1-\hat{p})\right]\end{aligned}$$

where we have used the fact that $\sum x_i = N\hat{p}$.

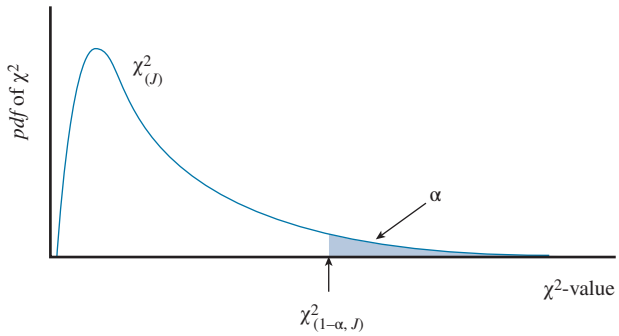


FIGURE C.15 **Critical value from a chi-square distribution.**

EXAMPLE C.21 | Likelihood Ratio Test of the Population Proportion

For our cereal box problem, $\hat{p} = 0.375$ and $N = 200$, so we have

$$\ln L(\hat{p}) = 200\left[0.375 \times \ln(0.375) + (1 - 0.375)\ln(1 - 0.375)\right]$$
$$= -132.3126$$

The value of the log-likelihood function assuming $H_0: p = 0.4$ is true is

$$\ln L(0.4) = \left(\sum_{i=1}^{N} x_i\right)\ln(0.4) + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1 - 0.4)$$
$$= 75 \times \ln(0.4) + (200 - 75) \times \ln(0.6)$$
$$= -132.5750$$

The problem is to assess whether -132.3126 is significantly different from -132.5750. The LR test statistic (C.25) is

$$LR = 2\left[\ln L(\hat{p}) - \ln L(0.4)\right]$$
$$= 2 \times (-132.3126 - (-132.575)) = 0.5247$$

If the null hypothesis $p = 0.4$ is true, then the LR test statistic has a $\chi^2_{(1)}$-distribution. If we choose $\alpha = 0.05$, then the test critical value is $\chi^2_{(0.95,1)} = 3.84$, the 95th percentile from the $\chi^2_{(1)}$-distribution. Since $0.5247 < 3.84$ we do not reject the null hypothesis.

The Wald Test In Figure C.14 it is clear that the distance (1/2)LR will depend on the curvature of the log-likelihood function. In Figure C.16 we show two log-likelihood functions with the hypothesized value c and the distances (1/2)LR for each of the log-likelihoods. The log-likelihoods have the same maximum value $\ln L\left(\hat{\theta}\right)$, but the values of the log-likelihood evaluated at the hypothesized value c are different.

The distance $\hat{\theta} - c$ translates into a larger value of (1/2)LR for the more highly curved log-likelihood, B, so it seems reasonable to construct a test measure by weighting the distance $\hat{\theta} - c$ by the magnitude of the log-likelihood's curvature, which we measure by the negative of its second derivative. This is exactly what the Wald statistic does:

$$W = \left(\hat{\theta} - c\right)^2 \left[-\frac{d^2 \ln L(\theta)}{d\theta^2}\right] \tag{C.26}$$

The value of the Wald statistic is larger for log-likelihood function B (more curved) than log-likelihood function A (less curved).

If the null hypothesis is true, then the Wald statistic (C.26) has a $\chi^2_{(1)}$ -distribution, and we reject the null hypothesis if $W \geq \chi^2_{(1-\alpha,1)}$. In more general situations we may test $J > 1$ hypotheses jointly, in which case we work with a chi-square distribution with J degrees of freedom, as shown in Figure C.15.

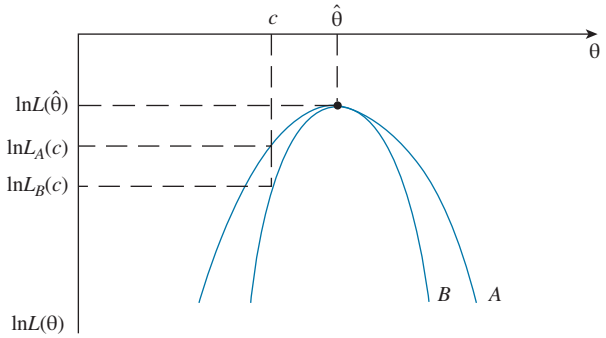


FIGURE C.16 **The Wald statistic.**

There is a linkage between the curvature of the log-likelihood function and the precision of maximum likelihood estimation. The greater the curvature of the log-likelihood function, the smaller the variance V in (C.23) and the more precise maximum likelihood estimation becomes, meaning that we have more **information** about the unknown parameter θ. Conversely, the more information we have about θ, the smaller the variance of the maximum likelihood estimator. Using this idea we define an **information measure** to be the reciprocal of the variance V,

$$I(\theta) = -E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right] = V^{-1} \tag{C.27}$$

As the notation indicates the information measure $I(\theta)$ is a function of the parameter θ. Substitute the information measure for the second derivative in the Wald statistic in (C.26) to obtain

$$W = \left(\hat{\theta} - c\right)^2 I(\theta) \tag{C.28}$$

In large samples the two versions of the Wald statistic are the same. An interesting connection here is obtained by rewriting (C.28) as

$$W = \left(\hat{\theta} - c\right)^2 V^{-1} = \left(\hat{\theta} - c\right)^2 \Big/ V \tag{C.29}$$

To implement the **Wald test**, we use the estimated variance

$$\hat{V} = \left[I\left(\hat{\theta}\right)\right]^{-1} \tag{C.30}$$

Then, taking the square root, we obtain the t-statistic in (C.22),

$$\sqrt{W} = \frac{\hat{\theta} - c}{\sqrt{\hat{V}}} = \frac{\hat{\theta} - c}{\text{se}\left(\hat{\theta}\right)} = t$$

That is, the t-test is also a Wald test.

EXAMPLE C.22 | Wald Test of the Population Proportion

In our blue box–green box example, we know that the maximum likelihood estimate $\hat{p} = 0.375$. To implement the Wald test we calculate

$$I(\hat{p}) = \hat{V}^{-1} = \frac{N}{\hat{p}(1-\hat{p})} = \frac{200}{0.375(1-0.375)} = 853.3333$$

where $V = p(1-p)/N$ and $\hat{V}$ were obtained in Section C.7.3. Then the calculated value of the Wald statistic is

$$W = (\hat{p} - c)^2 I(\hat{p}) = (0.375 - 0.4)^2 \times 853.3333 = 0.5333$$

In this case the value of the Wald statistic is close in magnitude to the LR statistic and the test conclusion is the same. Also, when testing one hypothesis, the Wald statistic is the square of the t-statistic, $W = t^2 = (-0.7303)^2 = 0.5333$.

The Lagrange Multiplier (LM) Test

The third testing procedure that comes from maximum likelihood theory is the Lagrange multiplier (LM) test. Figure C.17 illustrates another way to measure the distance between $\hat{\theta}$ and c. The slope of the log-likelihood function, which is sometimes called the *score*, is

$$s(\theta) = \frac{d \ln L(\theta)}{d\theta} \tag{C.31}$$

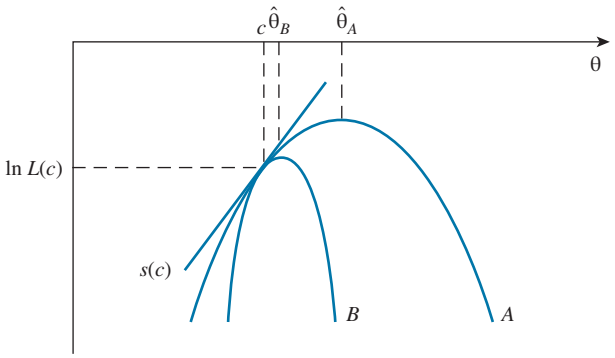


FIGURE C.17 **Motivating the Lagrange multiplier test.**

The slope of the log-likelihood function depends on the value of θ, as our function notation $s(\theta)$ indicates. The slope of the log-likelihood function at the maximizing value is zero, $s\left(\hat{\theta}\right) = 0$. The LM test examines the slope of the log-likelihood function at the point c. The logic of the test is that if $\hat{\theta}$ is close to c then the slope $s(c)$ of the log-likelihood function evaluated at c should be close to zero. In fact testing the null hypothesis $\theta = c$ is equivalent to testing $s(c) = 0$.

The difference between c and the maximum likelihood estimate $\hat{\theta}_B$ (maximizing ln L_B) is smaller than the difference between c and $\hat{\theta}_A$. In contrast to the Wald test, more curvature in the log-likelihood function implies a smaller difference between the maximum likelihood estimate and c. If we use the information measure $I(\theta)$ as our measure of curvature (more curvature means more information), the **Lagrange multiplier test** statistic can be written as

$$\text{LM} = \frac{\left[s(c)\right]^2}{I(\theta)} = \left[s(c)\right]^2\left[I(\theta)\right]^{-1} \tag{C.32}$$

The LM statistic for log-likelihood function A (less curved) is greater than the LM statistic for log-likelihood function B (more curved). If the null hypothesis is true, LM test statistic (C.32) has a $\chi^2_{(1)}$-distribution, and the rejection region is the same as for the LR and Wald tests. The LM, LR, and Wald tests are asymptotically equivalent and will lead to the same conclusion in sufficiently large samples.

In order to implement the LM test we can evaluate the information measure at the point $\theta = c$, so that it becomes

$$\text{LM} = \left[s(c)\right]^2\left[I(c)\right]^{-1}$$

In cases in which the maximum likelihood estimate is difficult to obtain (which it can be in more complex problems) the LM test has an advantage because $\hat{\theta}$ is not required. On the other hand, the Wald test in (C.28) uses the information measure evaluated at the maximum likelihood estimate $\hat{\theta}$,

$$W = \left(\hat{\theta} - c\right)^2 I\left(\hat{\theta}\right)$$

It is preferred when the maximum likelihood estimate and its variance are easily obtained. The likelihood ratio test statistic (C.25) requires calculation of the log-likelihood function at both the maximum likelihood estimate and the hypothesized value c. As noted, the three tests are asymptotically equivalent, and the choice of which to use is often made on the basis of convenience. In complex situations, however, the rule of convenience may not be a good one. The **likelihood ratio test** is relatively reliable in most circumstances, so if you are in doubt, it is a safe one to use.

EXAMPLE C.23 | Lagrange Multiplier Test of the Population Proportion

In the blue box–green box example, the value of the score, based on the first derivative shown just below (C.19), evaluated at the hypothesized value $c = 0.4$ is

$$s(0.4) = \frac{\sum x_i}{c} - \frac{N - \sum x_i}{1 - c} = \frac{75}{0.4} - \frac{200 - 75}{1 - 0.4} = -20.8333$$

The calculated information measure is

$$I(0.4) = \frac{N}{c(1 - c)} = \frac{200}{0.4(1 - 0.4)} = 833.3333$$

The value of the LM test statistic is

$$\text{LM} = \left[s(0.4)\right]^2 \left[I(0.4)\right]^{-1} = [-20.8333]^2 [833.3333]^{-1} = 0.5208$$

Thus, in our example, the values of the LR, Wald, and LM test statistics are very similar and give the same conclusion. This was to be expected, since the sample size $N = 200$ is large, and the problem is a simple one.

C.9 Algebraic Supplements

C.9.1 Derivation of Least Squares Estimator

In this section we illustrate how to use the least squares principle to obtain the sample mean as an estimator of the population mean. Represent a sample of data as $y_1, y_2, \ldots, y_N$. The population mean is $E(Y) = \mu$. The least squares principle says to find the value of μ that minimizes

$$S = \sum_{i=1}^{N} (y_i - \mu)^2$$

where S is the sum of squared deviations of the data values from μ.

The motivation for this approach can be deduced from the following example. Suppose you are going shopping at a number of shops along a certain street. Your plan is to shop at one store and return to your car to deposit your purchases. Then you visit a second store and return again to your car, and so on. After visiting each shop you return to your car. Where would you park to minimize the total amount of walking between your car and the shops you visit? You want to minimize the *distance* traveled. Think of the street along which you shop as a number line. The Euclidean distance between a shop located at y_i and your car at point μ is

$$d_i = \sqrt{(y_i - \mu)^2}$$

The squared distance, which is mathematically more convenient to work with, is

$$d_i^2 = (y_i - \mu)^2$$

To minimize the total squared distance between your parking spot μ and all the shops located at $y_1, y_2, \ldots, y_N$ you would minimize

$$S(\mu) = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_i - \mu)^2$$

which is the sum of squares function. Thus the least squares principle is really the least *squared distance* principle.

Since the values of y_i are known given the sample, the sum of squares function $S(\mu)$ is a function of the unknown parameter μ. Multiplying out the sum of squares, we have

$$S(\mu) = \sum_{i=1}^{N} y_i^2 - 2\mu \sum_{i=1}^{N} y_i + N\mu^2 = a_0 - 2a_1\mu + a_2\mu^2$$

EXAMPLE C.24 | Hip Data: Minimizing the Sum of Squares Function

For the data in Table C.1 we have

$$a_0 = \sum y_i^2 = 14880.1909,\ a_1 = \sum y_i = 857.9100,$$
$$a_2 = N = 50$$

The plot of the sum of squares parabola is shown in Figure C.18. The minimizing value appears to be a bit larger than 17 in the figure. Now we will determine the minimizing value exactly.

The value of μ that minimizes $S(\mu)$ is the "least squares estimate." From calculus, we know that the minimum of the function occurs where its slope is zero. See Appendix A.3.4. The function's derivative gives its slope, so by equating the first derivative of $S(\mu)$ to zero and solving, we can obtain the minimizing value exactly. The derivative of $S(\mu)$ is

$$\frac{dS(\mu)}{d\mu} = -2a_1 + 2a_2\mu$$

Setting the derivative to zero determines the least squares estimate of μ, which we denote as $\hat{\mu}$. Setting the derivative to zero,

$$-2a_1 + 2a_2\hat{\mu} = 0$$

Solving for $\hat{\mu}$ yields the formula for the least squares estimate,

$$\hat{\mu} = \frac{a_1}{a_2} = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}$$

Thus, the least squares estimate of the population mean is the sample mean, $\bar{y}$. This formula can be used in general, for any sample values that might be obtained, meaning that the least squares estimator is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} Y_i}{N} = \overline{Y}$$

For the hip data in Table C.1

$$\hat{\mu} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{857.9100}{50} = 17.1582$$

Thus, we estimate that the average hip size in the population is 17.1582 inches.

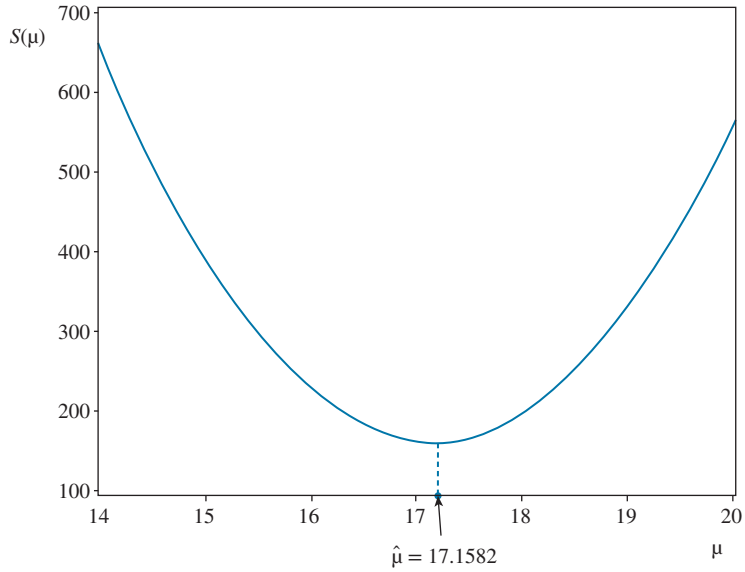


FIGURE C.18 **The sum of squares parabola for the hip data.**

C.9.2 Best Linear Unbiased Estimation

One of the powerful findings about the sample mean (which is also the least squares estimator) is that it is the best of all possible estimators that are both *linear* and *unbiased*. The fact that $\overline{Y}$ is the best linear unbiased estimator (BLUE) accounts for its wide use. In this context we mean by best that it is the estimator with the smallest variance of all linear and unbiased estimators. It is better to have an estimator with a smaller variance than one with a larger variance; it increases the

chances of getting an estimate close to the true population mean μ. This important result about the least squares estimator is true *if* the sample values $Y_i \sim (\mu, \sigma^2)$ are uncorrelated and identically distributed. It does not depend on the population being normally distributed. The fact that $\overline{Y}$ is BLUE is so important that we will prove it.

The sample mean is a weighted average of the sample values,

$$\begin{aligned}\overline{Y} &= \sum_{i=1}^{N} Y_i/N = \frac{1}{N}Y_1 + \frac{1}{N}Y_2 + \cdots + \frac{1}{N}Y_N \\ &= a_1Y_1 + a_2Y_2 + \cdots + a_NY_N \\ &= \sum_{i=1}^{N} a_iY\end{aligned}$$

where the weights $a_i = 1/N$. Weighted averages are also called linear combinations, so we call the sample mean a **linear estimator**. In fact, any estimator that can be written as $\sum_{i=1}^{N} a_iY_i$ is a linear estimator. For example, suppose the weights a_i^* are constants different from $a_i = 1/N$. Then we can define another linear estimator of μ as

$$\tilde{Y} = \sum_{i=1}^{N} a_i^*Y_i$$

To ensure that $\tilde{Y}$ is different from $\overline{Y}$, let us define

$$a_i^* = a_i + c_i = \frac{1}{N} + c_i$$

where c_i are constants that are not all zero. Thus,

$$\begin{aligned}\tilde{Y} &= \sum_{i=1}^{N} a_i^*Y_i = \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)Y_i \\ &= \sum_{i=1}^{N}\frac{1}{N}Y_i + \sum_{i=1}^{N} c_iY \\ &= \overline{Y} + \sum_{i=1}^{N} c_iY\end{aligned}$$

The expected value of the new estimator $\tilde{Y}$ is

$$\begin{aligned}E[\tilde{Y}] &= E\left[\overline{Y} + \sum_{i=1}^{N} c_iY_i\right] = \mu + \sum_{i=1}^{N} c_iE[Y_i] \\ &= \mu + \mu\sum_{i=1}^{N} c_i\end{aligned}$$

The estimator $\tilde{Y}$ is not unbiased unless $\sum c_i = 0$. We want to compare the sample mean to other linear and unbiased estimators, so we will assume that $\sum c_i = 0$ holds. Now we find the variance of $\tilde{Y}$. The linear unbiased estimator with the smaller variance will be best.

$$\begin{aligned}\text{var}(\tilde{Y}) &= \text{var}\left(\sum_{i=1}^{N} a_i^*Y_i\right) = \text{var}\left(\sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)Y_i\right) = \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2\text{var}(Y_i) \\ &= \sigma^2\sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2 = \sigma^2\sum_{i=1}^{N}\left(\frac{1}{N^2} + \frac{2}{N}c_i + c_i^2\right) = \sigma^2\left(\frac{1}{N} + \frac{2}{N}\sum_{i=1}^{N} c_i + \sum_{i=1}^{N} c_i^2\right) \\ &= \sigma^2/N + \sigma^2\sum_{i=1}^{N} c_i^2 \quad \left(\text{since}\sum_{i=1}^{N} c_i = 0\right) \\ &= \text{var}(\overline{Y}) + \sigma^2\sum_{i=1}^{N} c_i^2\end{aligned}$$

It follows that the variance of $\tilde{Y}$ must be greater than the variance of $\overline{Y}$, unless all the c_i values are zero, in which case $\tilde{Y} = \overline{Y}$.

C.10 Kernel Density Estimator

As econometricians, we work with data that are drawings from unknown distributions. For example, Figure C.19 shows the empirical distributions of two datasets, presented here as histograms. The variables X and Y are in the data file *kernel*. The problem before us is to estimate the density functions that yielded the observations. Knowledge about the distributions is important for statistical inference.

There are two main ways to estimate the distribution. We can use a parametric density estimator, or we can use a nonparametric **kernel density estimator**. In the **parametric approach**, we rely on density functions with well-defined functional forms characterized by parameters. For example, the normal probability density distribution $f(\cdot)$ has a specific functional form defined by two parameters—the mean μ and the standard deviation σ:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Once we have estimates of the mean and the standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, we plug these into the normal density function formula to obtain

$$\widehat{f(x)} = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)$$

Figure C.20 shows our application of this approach; the generated normal density functions are superimposed onto the histograms of the data. We have applied this parametric approach in the discussion about the Central Limit Theorem (C.3.4) and in discussion about ARCH models (Chapter 14).

The histogram of the variable X, on the left in Figure C.20, is unimodal, and the normal distribution appears to fit the shape of the data well. In contrast, the histogram of the variable Y on the right in Figure C.20 is bimodal, and the normal distribution is a poor representation of the underlying density function. We could try fitting the data with other parametric distributional forms, but rather than do that, let us adopt a nonparametric kernel density estimator to capture the shape of the data in a smooth continuous form.

Nonparametric methods do not require specific functional forms (e.g., the normal distribution formula) to generate the distribution. Instead, smoothing functions, called **kernels**, are used to "fit" the shape of the distribution of the data.

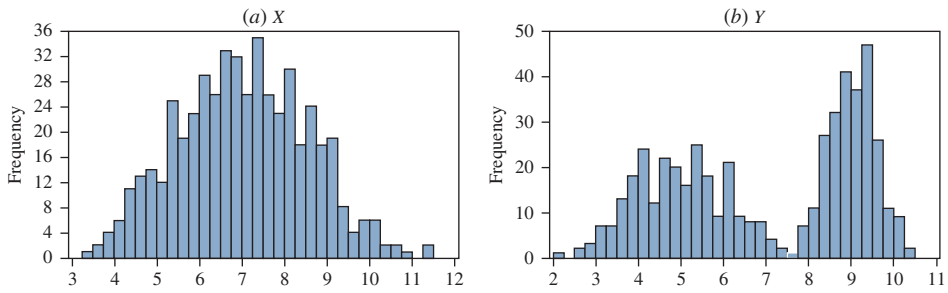


FIGURE C.19 **Histograms of variables (*a*) unimodal variable *X* and (*b*) bimodal variable *Y*.**

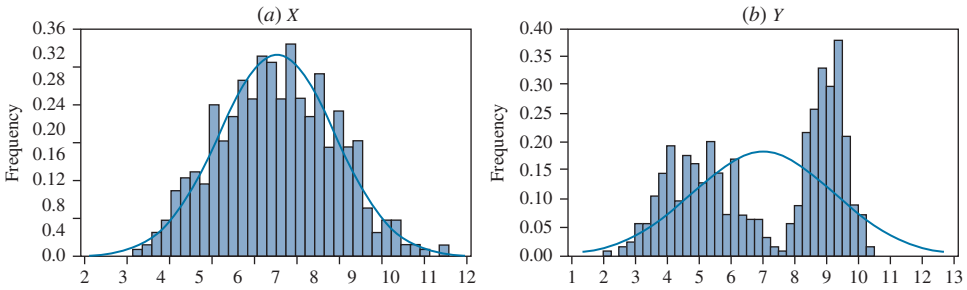


FIGURE C.20 **Parametric density estimator (*a*) unimodal variable *X* and (*b*) bimodal variable *Y*.**

The logic of the nonparametric approach can be grasped intuitively by thinking about how we set up histograms. Figure C.21 shows two histograms for the dataset Y. The one on the left has nine bins (i.e., the rectangles in the histogram) with bin width = 1 whereas the one on the right has many bins each with bin width = 0.1. The histogram with less bins has the higher frequency per bin as more observations fall into the larger bin width. More specifically, if x_k is the midpoint of the kth bin and h is the bin width, the range of values in the bin is $x_k \pm h/2$, and the frequency count n_k is the number of observations which falls in that range. The sum of all frequencies equals the sample size n, while the sum of the areas equals nh, since each area is $n_k h$ and $\sum_k n_k = n$. Note, too, that the shapes of the histograms are similar, but that the one with the larger bin width is "smoother" (fewer spikes and dips).

We can think of the histogram as a density function estimator $\widehat{f(x)}$, where x takes values over the domain of x and

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} 1\left(A_i\right)$$

The expression $1(A_i)$ is an **indicator function** taking on the value of 1 if A_i is true; A_i is the condition that x_i is in the same bin as x. For example, suppose we wish to find $\widehat{f(x)}$ for an x that lies in the kth bin. Then, A_i is true for all x_i such that $x_k - h/2 < x_i < x_k + h/2$. Thus, in the kth bin, $\sum_{i=1}^{n} 1\left(A_i\right) = n_k$, and the histogram density estimator for all x in the kth bin is $\widehat{f(x)} = n_k/nh$. The divisor nh ensures that the bin areas sum to one.

Now consider another density estimator where, instead of having a number of predetermined bins with midpoints x_k, we consider a bin with midpoint x and count the number of observations in the range $x \pm h/2$. If we repeat this process for all values of x, we can picture it as creating an infinite number of overlapping bins along the domain of x. In this case the density estimator is given by

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} 1\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right) = \frac{1}{nh}\sum_{i=1}^{n} 1\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right)$$

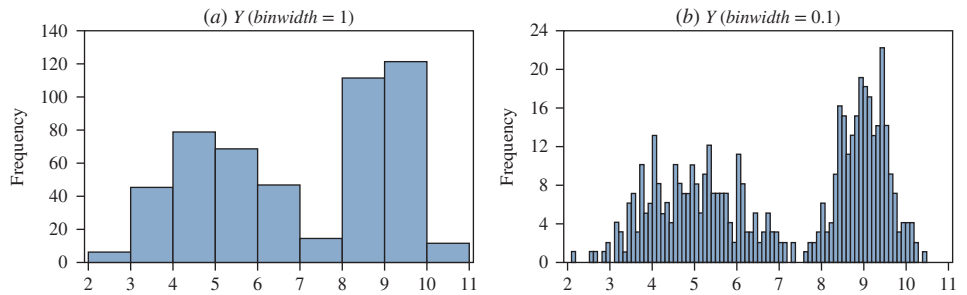


FIGURE C.21 **Histograms with different bin widths (*a*) width = 1 (*b*) width = 0.1.**

In practice, as you sum over the observations, the indicator function ensures that you only "count" the relevant observations. However, this density function will not be smooth, because each observation is given a weight of either zero or one—that is, it is either in or out, according to the condition specified in the indicator function.

Suppose we now replace this simple counting rule with a more sophisticated weighting function known as a **kernel**:

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where K is a kernel, h is a smoothing parameter called the **bandwidth**, and x is any value over the domain of possible values. There are many kernel functions; one of them is Gaussian and is described as follows:

$$K\left(\frac{x_i - x}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{h}\right)^2\right)$$

Figure C.22 shows the application of this kernel estimator to variable Y in data file *kernel* with four different bandwidths. Note how the shape of the density function is controlled by the bandwidth. The smaller the bandwidth, the better the fit, but there is a tradeoff between the number of "humps" captured and the smoothness of the fit. Intuitively, decreasing the bandwidth is like decreasing the bin width in the histogram, and the kernel is like a "counter"—but one which puts less weight on observations that are further away from the point being evaluated. (Imagine moving from the histogram on the right in Figure C.21 to the one on the left as you increase the bandwidth, and then imagine the use of the kernel to smooth the bars.) The kernel (Gaussian) density function with bandwidth equal to 0.4 appears to have captured the bimodality in the data.

There is a vast literature about the optimal choice of bandwidth as well as extensions of the nonparametric methods to regression analysis. Useful references include Pagan, A. and Ullah, A., *Nonparametric Econometrics*, Cambridge University Press, 1999; and Li, Q. and Racine, J.S. *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, 2007.

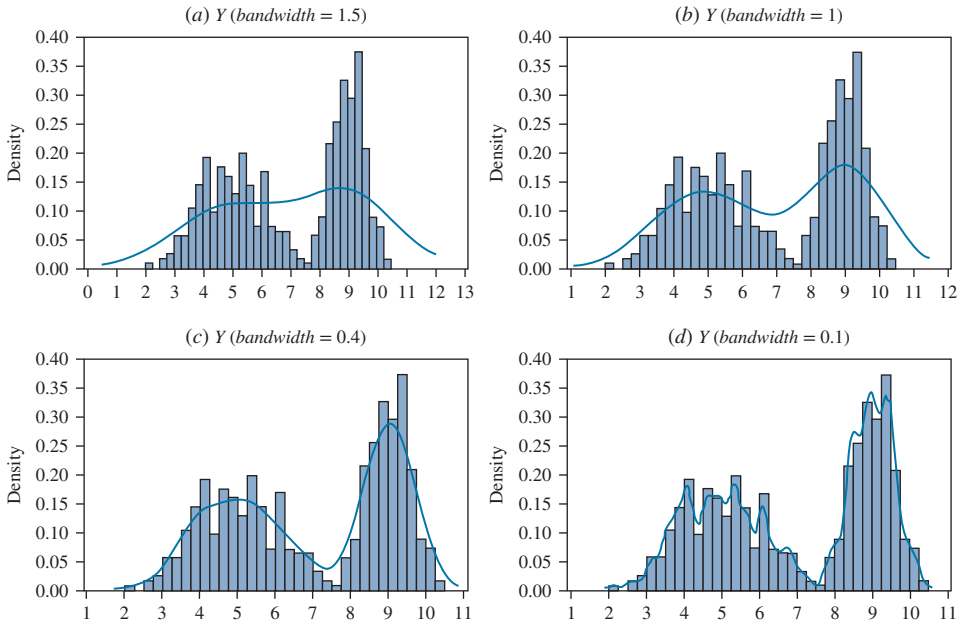


FIGURE C.22 **Fitting with a nonparametric density estimator (*a*) bandwidth = 1.5, (*b*) bandwidth = 1, (*c*) bandwidth = 0.4, and (*d*) bandwidth 0.1.**

C.11 Exercises

C.11.1 Problems

C.1 Suppose $Y_1, Y_2, \ldots, Y_N$ is a random sample from a population with mean μ and variance σ^2. Rather than using all N observations, consider an easy estimator of μ that uses only the first two observations

$$Y^* = \frac{Y_1 + Y_2}{2}$$

a. Show that Y^* is a linear estimator.
b. Show that Y^* is an unbiased estimator.
c. Find the variance of Y^*.
d. Explain why the sample mean of all N observations is a better estimator than Y^*.

C.2 Suppose that Y_1, Y_2, Y_3 is a random sample from a $N(\mu, \sigma^2)$ population. To estimate μ, consider the weighted estimator

$$\tilde{Y} = \frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3$$

a. Show that $\tilde{Y}$ is a linear estimator.
b. Show that $\tilde{Y}$ is an unbiased estimator.
c. Find the variance of $\tilde{Y}$ and compare it to the variance of the sample mean $\overline{Y}$.
d. Is $\tilde{Y}$ as good an estimator as $\overline{Y}$?
e. If $\sigma^2 = 9$, calculate the probability that each estimator is within one unit on either side of μ.

C.3 The hourly sales of fried chicken at Louisiana Fried Chicken are normally distributed with mean 2,000 pieces and standard deviation 500 pieces. What is the probability that in a 9-hour day more than 20,000 pieces will be sold?

C.4 Starting salaries for economics majors have a mean of \$47,000 and a standard deviation of \$8,000. What is the probability that a random sample of 40 economics majors will have an average salary of more than \$50,000?

C.5 A store manager designs a new accounting system that will be cost-effective if the mean monthly charge account balance is more than \$170. A sample of 400 accounts is randomly selected. The sample mean balance is \$178 and the sample standard deviation is \$65. Can the manager conclude that the new system will be cost-effective?

a. Carry out a hypothesis test to answer this question. Use the $\alpha = 0.05$ level of significance.
b. Compute the p-value of the test.

C.6 An econometric professor's rule of thumb is that students should expect to spend 2 hours outside of class on coursework for each hour in class. For a three-hour-per-week class, this means that students are expected to do 6 hours of work outside class. The professor randomly selects eight students from a class, and asks how many hours they studied econometrics during the past week. The sample values are 1, 3, 4, 4, 6, 6, 8, 12.

a. Assuming that the population is normally distributed, can the professor conclude at the 0.05 level of significance that the students are studying on average more than 6 hours per week?
b. Construct a 90% confidence interval for the population mean number of hours studied per week.

C.7 Modern labor practices attempt to keep labor costs low by hiring and laying off workers to meet demand. Newly hired workers are not as productive as experienced ones. Assume that assembly line workers with experience handle 500 pieces per day. A manager concludes that it is cost-effective to maintain the current practice if new hires, with a week of training, can process more than 450 pieces per day. A random sample of $N = 50$ trainees is observed. Let Y_i denote the number of pieces each handles on a randomly selected day. The sample mean is $\bar{y} = 460$, and the estimated sample standard deviation is $\hat{\sigma} = 38$.

a. Carry out a test of whether or not there is evidence to support the conjecture that current hiring procedures are effective, at the 5% level of significance. Pay careful attention when formulating the null and alternative hypotheses.

b. What exactly would a Type I error be in this example? Would it be a costly one to make?
c. Compute the p-value for this test.

C.8 To evaluate alternative retirement benefit packages for its employees, a large corporation must determine the mean age of its workforce. Assume that the age of its employees is normally distributed. Since the corporation has thousands of workers, a sample is to be taken. If the standard deviation of ages is known to be $\sigma = 21$ years, how large should the sample be to ensure that a 95% interval estimate of mean age is no more than four years wide?

C.9 Consider the discrete random variable Y that takes the values $y = 1, 2, 3,$ and 4 with probabilities 0.1, 0.2, 0.3, and 0.4, respectively.

a. Sketch this *pdf*.
b. Find the expected value of Y.
c. Find the variance of Y.
d. If we take a random sample of size $N = 3$ from this distribution, what are the mean and variance of the sample mean, $\overline{Y} = \left(y_1 + y_2 + y_3\right)/3$?

C.10 The sample proportion $\hat{p}$ is an estimator of the population proportion p. The variance of the estimator $\hat{p}$ is $\text{var}(\hat{p}) = p(1-p)/N$, where N is the sample size. Suppose we sample $N = 100$ voters. Of the 100 people sampled, 54 preferred candidate Hillary to candidate Donald.

a. Construct a 95% interval estimate of the population proportion using the approximately correct critical value 1.96 and the estimated variance $\widehat{\text{var}}(\hat{p}) = \hat{p}(1-\hat{p})/N$.
b. Calculate the alternative variance estimate, $\widetilde{\text{var}}(\hat{p}) = 0.5(1-0.5)/N$. Is this variance estimate larger or smaller than the one in part (a)? Will using the alternative variance make for a more conservative, wider, interval estimate or a less conservative, narrower, one?
c. Repeat the calculation of the interval estimate using the alternative variance estimate from part (b) and using the easier to work with critical value 2.0. Is it correct to say that this interval estimate has "a margin of error approximately equal to plus or minus 10 percent?"
d. Define the rough and conservative "margin of error" for the sample proportion interval to be $2\left[0.5(1-0.5)/N\right]^{1/2}$. Calculate the sample size required so that the margin of error is 0.07. What sample sizes are required for 0.05, 0.03, and 0.01 margins of error?
e. A February, 2017, Gallup poll on NAFTA (North American Free Trade Agreement) resulted in 48% saying it "has been a good thing." The poll was based on telephone interviews conducted Feb. 1-5, 2017, with a random sample of 1,035 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. Construct a conservative interval estimate of the true proportion of the 18 or older population thinking NAFTA has been a good thing. A news report based on the poll said "U.S. voters are deeply divided" on NAFTA. Do you think that is a fair statement? [One disheartening comment on the article by a reader said "I don't trust Poles."]

C.11 Let X denote the birthweight of a child, measured in hundreds of grams, whose mother did not smoke. Using a sample of $N = 968$ newly born children, we find the sample mean birthweight to be $\overline{X} = 34.2514$ hundred grams. Also $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = 33296.003$, $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^3 = -137910.04$, $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^4 = 6392783.3$

a. Use these values to compute the sample variance, as shown in (C.7) and the sample standard deviation, as shown in (C.9).
b. Use these values to compute $\tilde{\mu}_2$, $\tilde{\mu}_3$, $\tilde{\mu}_4$, as shown in Section C.4.2.
c. Calculate the skewness (S) and kurtosis (K) coefficients given in Section C.4.2. Are the values compatible with the normal distribution?
d. Test the normality of the data using the Jarque–Bera test in Section C.7.4.

C.12 Let Y denote the number of doctor visits in one month by a randomly chosen person. Assume that this count variable has a Poisson distribution with $E(Y) = \text{var}(Y) = \lambda$.

a. Calculate the probabilities $P(Y = 0)$, $P(Y = 1)$, and $P(Y = 2)$ assuming $\lambda = 1$.
b. We choose a random sample of $N = 3$ individuals and observe that the first and second people had two doctor visits, and the third person had one. Calculate the joint probability $P\left(Y_1 = 2, Y_2 = 2, Y_3 = 1\right)$ given that $\lambda = 1$.
c. Show that in general $P\left(Y_1 = 2, Y_2 = 2, Y_3 = 1 \middle| \lambda\right) = 0.25\lambda^5 e^{-3\lambda}$.

d. The likelihood function is $L(\lambda|Y_1 = 2, Y_2 = 2, Y_3 = 1) = 0.25\lambda^5 e^{-3\lambda}$. Write down the algebraic form of the log-likelihood, $\ln L(\lambda|Y_1 = 2, Y_2 = 2, Y_3 = 1)$.

e. Find the first derivative of the log-likelihood, set it to zero, and solve for the solution value $\tilde{\lambda}$.

f. Find the second derivative of the log-likelihood. Determine the sign of this derivative.

g. Can we claim that $\tilde{\lambda}$ is the maximum likelihood estimate of $E(Y) = \text{var}(Y) = \lambda$?

C.13 This exercise extends Exercise C.12 to the general case with a random sample of N observations, $Y_1, \ldots, Y_N$ from a population. Each outcome is assumed to have a Poisson distribution with $E(Y) = \text{var}(Y) = \lambda$.

a. Show that the log-likelihood function is $\ln L(\lambda|y_1, \ldots, y_N) = (\ln \lambda) \sum_{i=1}^N y_i - N\lambda - \sum_{i=1}^N \ln(y_i!)$.

b. Show that the maximum likelihood estimate is $\tilde{\lambda} = \sum_{i=1}^N y_i / N$.

c. Show that the second derivative of the log-likelihood function is $-\left(\sum_{i=1}^N y_i\right)/\lambda^2$. What is the sign of the second derivative?

d. The maximum likelihood estimator is $\tilde{\lambda} = \sum_{i=1}^N Y_i / N$. Assuming we have a random sample from a population with $E(Y) = \text{var}(Y) = \lambda$, find $E(\tilde{\lambda})$ and $\text{var}(\tilde{\lambda})$.

e. The information measure $I(\lambda) = -\left\{E\left[\frac{d^2 \ln L(\lambda)}{d\lambda^2}\right]\right\}$, where $\left[\frac{d^2 \ln L(\lambda)}{d\lambda^2}\right] = -\left(\sum_{i=1}^N Y_i\right)/\lambda^2$. Show that the information measure in this case is $I(\lambda) = N/\lambda$.

C.14 Let X denote the birthweight of a child, measured in hundreds of grams. Consider children whose mothers smoked ($SMOKE = 1$) and children whose mothers did not smoke ($SMOKE = 0$). Summary statistics for the birthweights for these two groups are in Table C.6.

TABLE C.6 Summary Statistics for Birthweights

SMOKE	*N*	Mean	Variance	Std. Dev.	Skewness	Kurtosis
0	968	34.25	34.43	5.87	−0.71	5.58
1	232	31.37	34.42	5.87	−1.26	7.66

a. Use the Jarque–Bera test to test the normality of each of these populations. Do we reject the null hypothesis of normality or fail to reject normality?

b. Construct a 95% interval estimate for the population mean birthweight born to mothers who did not smoke, μ_0. Construct a 95% interval estimate for the population mean birthweight born to mothers who did smoke, μ_1. Select any value c in the 95% interval estimate for μ_0. What is the outcome of a two-tail test of the hypothesis $\mu_1 = c$ using the 5% level of significance?

c. Test the null hypothesis that the population mean birthweight is the same for the two populations, $H_0: \mu_0 = \mu_1$ against the alternative $H_1: \mu_0 \neq \mu_1$. Explain your choice to use a pooled variance estimator or to assume that the pooled variance is inappropriate. Use the 5% level of significance.

d. Repeat the test in part (c) for the null and alternative hypotheses $H_0: \mu_0 \leq \mu_1$ and $H_1: \mu_0 > \mu_1$.

C.15 In this exercise we use the data from Exercise C.12 and the results in Exercise C.13 to carry out a hypothesis test concerning the parameter λ in the Poisson distribution.

a. Using the maximum likelihood estimate from Exercise C.12, compute the information measure $I(\tilde{\lambda})$, given in Exercise C.13 (e).

b. Carry out a likelihood ratio test of the null hypothesis $H_0: \lambda = 1$, using the test statistic in equation (C.25), versus the alternative $H_1: \lambda \neq 1$ at the 5% level of significance.

c. Use the Wald statistic in equation (C.26) to carry out the test from part (b).

d. An alternative version of the Wald statistic replaces the second derivative term, $-d^2 \ln L(\lambda)/d\lambda^2$, with $I(\tilde{\lambda})$, as shown in equation (C.28). Carry out the test from part (b) using the modified Wald test.

e. Evaluate the score function, shown in equation (C.31), assuming the null hypothesis is true.

f. Evaluate the information measure $I(\lambda)$ assuming the null hypothesis is true.

g. Using the results in parts (e) and (f), carry out the LM test of the null hypothesis in part (b).

C.16 Two independent food scientists are researching the shelf-life (Y) of "Bill's Big Red" spaghetti sauce. The first collects a random sample of 25 jars and finds their average shelf life to be $\overline{Y} = 48$ months. The second researcher collects a random sample of 100 jars and finds their average shelf life to be $\overline{Y}_2 = 40$ months.

a. Find the ratio of the standard error of $\overline{Y}_1$ relative to the standard error of $\overline{Y}_2$.

b. A combined estimate can be obtained by finding the weighted average $\tilde{Y} = c\overline{Y}_1 + (1-c)\overline{Y}_2$. Is there any value of c that makes this estimator of μ unbiased?

c. What value of c yields the combined estimate with the smallest standard error? Explain the intuition behind your solution, and why weighting the two means equally, with $c = 0.5$, is not the best choice.

C.17 Suppose school children are subjected to a standardized math test each spring. In the population of comparable children, the test score Y is normally distributed with mean 500 and standard deviation 100, $Y \sim N(\mu = 500, \sigma^2 = 100^2)$. It is claimed that reducing class sample size will increase test scores.

a. How can we tell if reducing class size actually does increase test scores? Would you be convinced if a sample of $N = 25$ students from the smaller classes had an average test score of 510? Calculate the probability of obtaining a sample mean of $\overline{Y} = 510$, or more, even if smaller classes actually have no effect on test performance.

b. Show that a class average of 533 will be reached by chance only 5% of the time, if the smaller class sizes have no effect. Is the following statement correct or incorrect? "We can conclude that smaller classes raise average test scores if a class of 25 students has an average test score of 533 or better, with this result being due to sampling error with probability 5%."

c. Suppose that smaller classes actually do improve the average mean population test score to 550. What is the probability of observing a class of 25 with an average score of 533 or better? If our objective is to determine whether smaller classes increase test scores, is it better for this number to be larger or smaller?

d. If smaller classes increase average test score to 550, what is the probability of having a small class average of less than or equal to 533?

e. Draw a figure showing two normal distributions, one with mean 500 and standard deviation 100, and the other with mean 550 and standard deviation 100. On the figure locate the value 533. In part (b) we showed that if the change in class size has no effect on test scores, we would still obtain a class average of 533 or more by chance 5% of the time; we would incorrectly conclude that the smaller classes helped test scores, which is a Type I error. In part (d) we derived the probability that we would obtain a class average test score of less than 533, making us unable to conclude that smaller classes help, even though smaller classes did help. This is a Type II error. If we push the threshold to the right, say 540, what happens to Type I and Type II errors? If we push the threshold to the left, say 530, what happens to the probability of Type I and Type II errors?

C.11.2 Computer Exercises

C.18 Does being in a small class help primary school students learning, and performance on achievement tests? Use the sample data file *star5_small* to explore this question.

a. Consider students in regular-sized classes, with *REGULAR* = 1. Construct a histogram of *MATHSCORE*. Carry out the Jarque–Bera test for normality at the 5% level of significance. What do you conclude about the normality of the data?

b. Calculate the sample mean, standard deviation and standard error of the mean for *MATHSCORE* in regular-sized classes. Use the t-statistic in equation (C.16) to test the null hypothesis that the population mean (the population of students who are enrolled in regular-sized classes) μ_R is 490 versus the alternative that it isn't. Use the 5% level of significance. What is your conclusion?

c. Given the result of the normality test in (a), do you think the test in part (b) is justifiable? Explain your reasoning.

d. Construct a 95% interval estimate for the mean μ_R.

e. Repeat the test in (b) for the population of students in small classes, *SMALL* = 1. Denote the population mean for these students as μ_S. Use the 5% level of significance. What is your conclusion?

f. Let μ_R and μ_S denote the population mean test scores on the math achievement test, *MATHSCORE*. Using the appropriate test, outlined in Section C.7.2, test the null hypothesis $H_0: \mu_S - \mu_R \leq 0$ against the alternative $H_1: \mu_S - \mu_R > 0$. Use the 1% level of significance. Does it appear that being in a small class increases the expected math test score, or not?

C.19 Does having a household member with an advanced degree increase household income relative to a household that includes a member having only a college degree? Use the sample data file *cex5_small* to explore this question.

- **a.** Construct a histogram of incomes for households that include a member with an advanced degree. Construct another histogram of incomes for households that include a member with a college degree. What do you observe about the shape and location of these two histograms?
- **b.** In the sample that includes a member with an advanced degree, what percentage of households have household incomes greater than \$10,000 per month? What is the percentage for households that include a member having a college degree?
- **c.** Test the null hypothesis that the population mean income for households including a member with an advanced degree, μ_{ADV}, is less than, or equal to, \$9,000 per month against the alternative that it is greater than \$9,000 per month. Use the 5% level of significance.
- **d.** Test the null hypothesis that the population mean income for households including a member with a college degree, μ_{COLL}, is less than, or equal to, \$9,000 per month against the alternative that it is greater than \$9,000 per month. Use the 5% level of significance.
- **e.** Construct 95% interval estimates for μ_{ADV} and μ_{COLL}.
- **f.** Test the null hypothesis $\mu_{ADV} \leq \mu_{COLL}$ against the alternative $\mu_{ADV} > \mu_{COLL}$. Use the 5% level of significance. What is your conclusion?

C.20 How much variation is there in household incomes in households including a member with an advanced degree? Use the sample data file *cex5_small* to explore this question. Let σ^2_{ADV} denote the population variance.

- **a.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} > 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region. What is the ***p*-value** for this test?
- **b.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} < 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region. What is the *p*-value for this test?
- **c.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} \neq 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region.

C.21 School officials consider performance on a standardized math test acceptable if 40% of the population of students score at least 500 points. Use the sample data file *star5_small* to explore this topic.

- **a.** Compute the sample proportion of students enrolled in regular-sized classes who score 500 points or more. Calculate a 95% interval estimate of the population proportion. Based on this interval can we reject the null hypothesis that the population proportion of students in regular-sized classes who score 500 points or better is $p = 0.4$?
- **b.** Test the null hypothesis that the population proportion p of students in a regular-sized class who score 500 points or more is less than or equal to 0.4 against the alternative that the true proportion is greater than 0.4. Use the 5% level of significance.
- **c.** Test the null hypothesis that the population proportion p of students in a regular-sized class who score 500 points or more is equal to 0.4 against the alternative that the true proportion is less than 0.4. Use the 5% level of significance.
- **d.** Repeat parts (a)–(c) for students in small classes.

C.22 Consider two populations of Chinese chemical firms: those who export their products and those who do not. Let us consider the sales revenue for these two types of firms. Use the data file *chemical_small* for this exercise. It contains data on 1200 firms in 2006.

- **a.** The variable *LSALES* is ln(*SALES*). Construct a histogram for this variable and test whether the data are normally distributed using the Jarque–Bera test with 10% level of significance.
- **b.** Create the variable $SALES = \exp(LSALES)$. Construct a histogram for this variable and test whether the data are normally distributed using the Jarque–Bera test with 10% level of significance.
- **c.** Consider two populations of firms: those who export ($EXPORT = 1$) and those who do not ($EXPORT = 0$). Let μ_1 be the population mean of *LSALES* for firms that export, and let μ_0 be the population mean of *LSALES* for firms that do not export. Estimate the difference in means $\mu_1 - \mu_0$ and interpret this value. [*Hint*: Use the properties of differences in log-variables.]
- **d.** Test the hypothesis that the means of these two populations are equal. Use the test that assumes the population variances are unequal. What do you conclude?

C.23 Does additional education have as large a payoff for females as males? Use the data file *cps5* to explore this question. If your software does not permit using this larger sample use *cps5_small*.

- **a.** Calculate the sample mean wage of females who have 12 years of education. Calculate the sample mean wage of females with 16 years of education. What did you discover?
- **b.** Calculate a 95% interval estimate for the population mean wage of females with 12 years of education. Repeat the calculation for the wages of females with 16 years of education. Do the intervals overlap?
- **c.** Calculate the sample mean wage of males who have 12 years of education. Calculate the sample mean wage of males with 16 years of education. What did you discover? How does the difference in wages for males compare to the difference of wages for females in part (a)?
- **d.** Calculate a 95% interval estimate for the population mean wage of males with 12 years of education. Repeat the calculation for the wages of males with 16 years of education. Does the interval for males with 12 years of education overlap with the comparable interval for females? Does the interval for males with 16 years of education overlap with the comparable interval for females?
- **e.** Denote the population means of interest by μ_{F16}, μ_{F12}, μ_{M16}, μ_{M12} where F and M denote female and male, and 12 and 16 denote years of education. Estimate the parameter $\theta = (\mu_{F16} - \mu_{F12}) - (\mu_{M16} - \mu_{M12})$ by replacing population means by sample means.
- **f.** Calculate a 95% interval estimate of θ. Based on the interval estimate, what can you say about the benefits of the addition of four years of education for males versus females? Use the 97.5 percentile from the standard normal, 1.96, when calculating the interval estimate.

C.24 How much does the variation in wages change when individuals receive more education? Is the variation different for males and females? Use the data file *cps5* to explore this question. If your software does not permit using this larger sample use *cps5_small*.

- **a.** Calculate the sample variance of wages of females who have 12 years of education. Calculate the sample variance of wages of females who have 18 years of education. What did you discover?
- **b.** Carry out a two-tail test, using a 5% level of significance, of the hypothesis that the variance of wage is the same for females with 12 years of education and females with 18 years of education.
- **c.** Calculate the sample variance of wages of males who have 12 years of education. Calculate the sample variance of wages of males who have 18 years of education. What did you discover?
- **d.** Carry out a two-tail test, using a 5% level of significance, of the hypothesis that the variance of wage is the same for males with 12 years of education and males with 18 years of education.
- **e.** Carry out a two-tail test of the null hypothesis that the mean wage for males with 18 years of education is the same as the mean wage of females with 18 years of education. Use the 1% level of significance.

C.25 What happens to the household budget share of necessity items, like food, when total household expenditures increase? Use data file *malwai_small* for this exercise.

- **a.** Obtain the summary statistics, including the median and 90th percentile, of total household expenditures.
- **b.** Construct a 95% interval estimate for the proportion of income spent on food by households with total expenditures less than or equal to the median.
- **c.** Construct a 95% interval estimate for the proportion of income spent on food by households with total expenditures more than or equal to the 90th percentile.
- **d.** Summarize your findings from parts (b) and (c).
- **e.** Test the null hypothesis that the population mean proportion of income spent on food by households is 0.4. Use a two-tail test and the 5% level of significance. Carry out the test separately using the complete sample, and using the samples of households with total expenditures less than or equal to the median, and again for households whose total expenditures are in the top 10%.

C.26 At the famous Fulton Fish Market in New York City sales of Whiting (a type of fish) vary from day to day. Over a period of several months, daily quantities sold (in pounds) were observed. These data are in the data file *fultonfish*.

- **a.** Using the data for Monday sales, test the null hypothesis that the mean quantity sold is greater than or equal to 10,000 pounds a day, against the alternative that the mean quantity sold is less than 10,000 pounds. Use the $\alpha = 0.05$ level of significance. Be sure to (i) state the null and alternative hypotheses, (ii) give the test statistic and its distribution, (iii) indicate the rejection region, including

a sketch, (iv) state your conclusion, and (v) calculate the *p*-value for the test. Include a sketch showing the *p*-value.

b. Assume that daily sales on Tuesday (X_2) and Wednesday (X_3) are normally distributed with means μ_2 and μ_3, and variances σ_2^2 and σ_3^2, respectively. Assume that sales on Tuesday and Wednesday are independent of each other. Test the hypothesis that the variances σ_2^2 and σ_3^2 are equal against the alternative that the variance on Tuesday is larger. Use the $\alpha = 0.05$ level of significance. Be sure to (i) state the null and alternative hypotheses, (ii) give the test statistic and its distribution, (iii) indicate the rejection region, including a sketch, (iv) state your conclusion, and (v) calculate the *p*-value for the test. Include a sketch showing the *p*-value.

c. We wish to test the hypothesis that mean daily sales on Tuesday and Wednesday are equal against the alternative that they are not equal. Using the result in part (b) as a guide to the appropriate version of the test (Appendix C.7), carry out this hypothesis test using the 5% level of significance.

d. Let the daily sales for Monday, Tuesday, Wednesday, Thursday, and Friday be denoted as X_1, X_2, X_3, X_4, and X_5, respectively. Assume that $X_i \sim NID(\mu_i, \sigma_i^2)$. Define total weekly sales as $W = X_1 + X_2 + X_3 + X_4 + X_5$. Derive the expected value and variance of W, using appropriate theorems about normal distributions. Be sure to show your work and justify your answer.

e. Referring to part (d), let $E(W) = \mu$. Assume that we estimate μ using

$$\hat{\mu} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \overline{X}_4 + \overline{X}_5$$

where $\overline{X}_i$ is the sample mean for the *i*th day. Derive the probability distribution of $\hat{\mu}$ and construct an approximate (valid in large samples) 95% interval estimate for μ. Justify the validity of your interval estimator.

C.27 A **credit score** is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of the person. A credit score is primarily based on a credit report information typically sourced from credit bureaus. Use the data file *lasvegas* for this exercise.

a. Construct a histogram for the variable *CREDIT*. Does the histogram look symmetrical and "bell-shaped?" Test the normality of the variable *CREDIT* using the Jarque–Bera test and level of significance 5%.

b. Let two populations of *CREDIT* be defined by those who were delinquent (*DELINQUENT* = 1) and those who were not delinquent (*DELINQUENT* = 0). Using the test described in Section C.7.3, carry out a test of the hypothesis that the variances in these two populations are equal against the alternative that they are not equal. Use the 5% level of significance.

c. Use the appropriate one-tail test in Section C.7.2, based on your answer in part (b), to test the equality of *CREDIT* means for the two populations.

d. Using the test in Section C.7.1, test the null hypothesis that the variance of the population who was not delinquent is 3600 versus the alternative that it is not 3600.

C.28 Is it true that more capable individuals ultimately attain more years of schooling? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979.

a. In the data the variable *SCORE* is an index based on 10 aptitude/IQ tests given in 1980. We can loosely use this variable as some measure of ability. Construct a histogram of *SCORE*. What are the sample mean and the standard deviation?

b. The variable *EDUC* is the individual's years of schooling completed by 1993. What percentage of the men had completed at least 12 years of education by 1993?

c. Calculate the sample mean number of years of schooling completed by men with *SCORE* greater than or equal to zero. Calculate the sample mean number of years of schooling completed by men with *SCORE* less than zero. Test the null hypothesis that the population of men with $SCORE \geq 0$ have mean years of education, μ_1, that is greater than the mean number of years of education, μ_0, for those with lower scores. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

d. Some of the men came from broken homes, as indicated by the variable *BROKEN*. Test the null hypothesis that the population of men from broken homes have mean years of education, μ_1, that is less than the mean number of years of education, μ_0, for those who were not from broken homes. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

C.29 Do more highly educated parents tend to have more educated children? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979.

a. The variable *EDUC* is the individual's years of schooling completed by 1993. What percentage of the men had completed at least 16 years of education by 1993? What percentage of the men's mothers had at least 16 years of education? What percentage of fathers had at least 16 years of education?

b. Calculate the sample mean number of years of schooling completed by men with fathers who had 16 or more years of education. Calculate the sample mean number of years of schooling completed by men with fathers who had less than 16 years of education. Test the null hypothesis that the population of men with more educated fathers have mean years of education, μ_1, that is greater than the mean number of years of education, μ_0, for those with less educated fathers. State the null and alternative hypothesis, give the test statistic, and your conclusion using a 5% level of significance.

c. Investigate the question of whether more highly educated men, those with more than 12 years of schooling, tend to marry more highly educated women, those with more than 12 years of schooling. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

C.30 Do households with more children tend to result in more broken homes? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979. It includes the number of siblings the man had as well as whether he came from a broken home.

a. Create the variable $KIDS = SIBS + 1$. To simplify the following arithmetic, let $KIDS = 3$ if the number of household children is equal to 3 or more. The variable *KIDS* takes the values 1, 2, and 3. Calculate the number of men who came from families with $KIDS = 1$, and $KIDS = 2$, and $KIDS = 3$.

b. Calculate the number of households that were broken having 1, 2, or 3 children. Calculate the number of households that were not broken with $KIDS = 1$, $KIDS = 2$, and $KIDS = 3$.

c. The famous statistician Karl Pearson developed a test for the null hypothesis that two characteristics are unrelated versus the alternative that they are related. If the number of children and broken homes are unrelated, we should expect 176.167 of 1057 households with each of the six possible outcomes. Pearson's chi-square test is calculated as

$$PEARSON = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_6 - E_6)^2}{E_6}$$

where E_i is the "expected" number of outcomes and O_i is the "observed" number of outcomes for each of six outcomes. If there is no relation between the variables the test statistic has a $\chi^2_{(m)}$ distribution, with $m = (c_1 - 1) \times (c_2 - 1)$ degrees of freedom, where c_1 is the number of categories for variable 1 and c_2 is the number of categories for variable 2. The null hypothesis that the variables are unrelated is rejected if the value of *PEARSON* is greater than the $100(1-\alpha)$-percentile from the chi-square distribution. Carry out Pearson's test for the existence of a relationship between *BROKEN* and *KIDS* at the 5% level.

d. Explore your software. Does it have a command to automatically create two-way tables of frequencies? Does it have a command to calculate Pearson's chi-square statistic? If so, carry out the test in part (c) *without* modifying the variable *KIDS* to have only three outcomes. Report the two-way table and the test result.

Appendix D

Statistical Tables

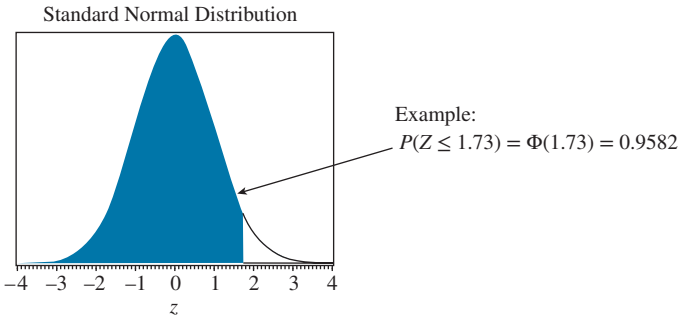


TABLE D.1 **Cumulative Probabilities for the Standard Normal Distribution $\Phi(z) = P(Z \leq z)$**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Source: This table was generated using the SAS® function PROBNORM.

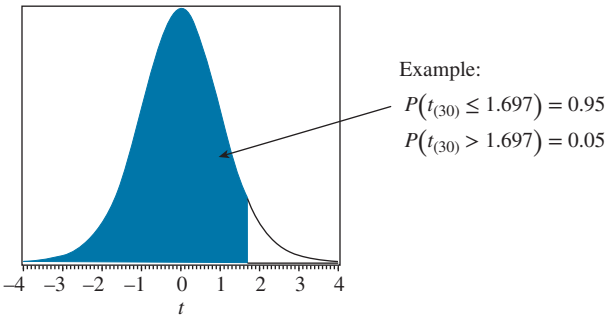


TABLE D.2 **Percentiles of the *t*-distribution**

df	$t_{(0.90, df)}$	$t_{(0.95, df)}$	$t_{(0.975, df)}$	$t_{(0.99, df)}$	$t_{(0.995, df)}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
36	1.306	1.688	2.028	2.434	2.719
37	1.305	1.687	2.026	2.431	2.715
38	1.304	1.686	2.024	2.429	2.712
39	1.304	1.685	2.023	2.426	2.708
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
∞	1.282	1.645	1.960	2.326	2.576

Source: This table was generated using the SAS® function TINV.

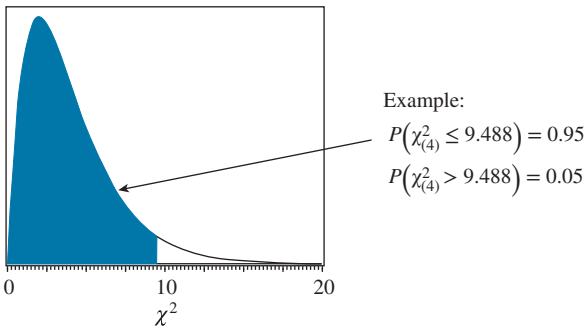
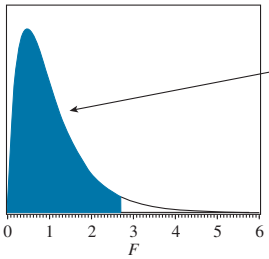


TABLE D.3 **Percentiles of the Chi-square Distribution**

df	$\chi^2_{(0.90,\,df)}$	$\chi^2_{(0.95,\,df)}$	$\chi^2_{(0.975,\,df)}$	$\chi^2_{(0.99,\,df)}$	$\chi^2_{(0.995,\,df)}$
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672
35	46.059	49.802	53.203	57.342	60.275
40	51.805	55.758	59.342	63.691	66.766
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.379	91.952
70	85.527	90.531	95.023	100.425	104.215
80	96.578	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169
110	129.385	135.480	140.917	147.414	151.948
120	140.233	146.567	152.211	158.950	163.648

Source: This table was generated using the SAS® function CINV.



Example:

$P(F_{(4,30)} \leq 2.69) = 0.95$

$P(F_{(4,30)} > 2.69) = 0.05$

TABLE D.4 **95th Percentile for the *F*-distribution**

v_2/v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.10	252.20	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.62	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.82	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.68	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.64	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.60	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.58	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.00

Source: This table was generated using the SAS® function FINV.

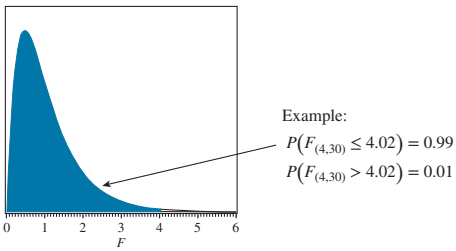


TABLE D.5 **99th Percentile for the *F*-distribution**

ν_2/ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6157.28	6208.73	6260.65	6313.03	6365.87
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.48	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.32	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.65	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.20	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	3.91
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.87
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.42
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.36	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.01
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.60	2.44	2.28	2.10	1.89
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.02	1.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.61	2.46	2.31	2.14	1.96	1.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.91	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.00

Source: This table was generated using the SAS® function FINV.

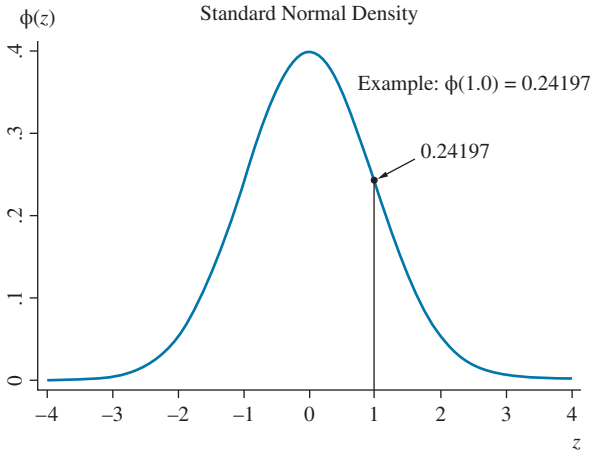


TABLE D.6 **Standard Normal pdf Values $\phi(z)$**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.39894	0.39892	0.39886	0.39876	0.39862	0.39844	0.39822	0.39797	0.39767	0.39733
0.1	0.39695	0.39654	0.39608	0.39559	0.39505	0.39448	0.39387	0.39322	0.39253	0.39181
0.2	0.39104	0.39024	0.38940	0.38853	0.38762	0.38667	0.38568	0.38466	0.38361	0.38251
0.3	0.38139	0.38023	0.37903	0.37780	0.37654	0.37524	0.37391	0.37255	0.37115	0.36973
0.4	0.36827	0.36678	0.36526	0.36371	0.36213	0.36053	0.35889	0.35723	0.35553	0.35381
0.5	0.35207	0.35029	0.34849	0.34667	0.34482	0.34294	0.34105	0.33912	0.33718	0.33521
0.6	0.33322	0.33121	0.32918	0.32713	0.32506	0.32297	0.32086	0.31874	0.31659	0.31443
0.7	0.31225	0.31006	0.30785	0.30563	0.30339	0.30114	0.29887	0.29659	0.29431	0.29200
0.8	0.28969	0.28737	0.28504	0.28269	0.28034	0.27798	0.27562	0.27324	0.27086	0.26848
0.9	0.26609	0.26369	0.26129	0.25888	0.25647	0.25406	0.25164	0.24923	0.24681	0.24439
1.0	0.24197	0.23955	0.23713	0.23471	0.23230	0.22988	0.22747	0.22506	0.22265	0.22025
1.1	0.21785	0.21546	0.21307	0.21069	0.20831	0.20594	0.20357	0.20121	0.19886	0.19652
1.2	0.19419	0.19186	0.18954	0.18724	0.18494	0.18265	0.18037	0.17810	0.17585	0.17360
1.3	0.17137	0.16915	0.16694	0.16474	0.16256	0.16038	0.15822	0.15608	0.15395	0.15183
1.4	0.14973	0.14764	0.14556	0.14350	0.14146	0.13943	0.13742	0.13542	0.13344	0.13147
1.5	0.12952	0.12758	0.12566	0.12376	0.12188	0.12001	0.11816	0.11632	0.11450	0.11270
1.6	0.11092	0.10915	0.10741	0.10567	0.10396	0.10226	0.10059	0.09893	0.09728	0.09566
1.7	0.09405	0.09246	0.09089	0.08933	0.08780	0.08628	0.08478	0.08329	0.08183	0.08038
1.8	0.07895	0.07754	0.07614	0.07477	0.07341	0.07206	0.07074	0.06943	0.06814	0.06687
1.9	0.06562	0.06438	0.06316	0.06195	0.06077	0.05959	0.05844	0.05730	0.05618	0.05508
2.0	0.05399	0.05292	0.05186	0.05082	0.04980	0.04879	0.04780	0.04682	0.04586	0.04491
2.1	0.04398	0.04307	0.04217	0.04128	0.04041	0.03955	0.03871	0.03788	0.03706	0.03626
2.2	0.03547	0.03470	0.03394	0.03319	0.03246	0.03174	0.03103	0.03034	0.02965	0.02898
2.3	0.02833	0.02768	0.02705	0.02643	0.02582	0.02522	0.02463	0.02406	0.02349	0.02294
2.4	0.02239	0.02186	0.02134	0.02083	0.02033	0.01984	0.01936	0.01888	0.01842	0.01797
2.5	0.01753	0.01709	0.01667	0.01625	0.01585	0.01545	0.01506	0.01468	0.01431	0.01394
2.6	0.01358	0.01323	0.01289	0.01256	0.01223	0.01191	0.01160	0.01130	0.01100	0.01071
2.7	0.01042	0.01014	0.00987	0.00961	0.00935	0.00909	0.00885	0.00861	0.00837	0.00814
2.8	0.00792	0.00770	0.00748	0.00727	0.00707	0.00687	0.00668	0.00649	0.00631	0.00613
2.9	0.00595	0.00578	0.00562	0.00545	0.00530	0.00514	0.00499	0.00485	0.00470	0.00457
3.0	0.00443	0.00430	0.00417	0.00405	0.00393	0.00381	0.00370	0.00358	0.00348	0.00337

Source: This table was generated using the SAS® function PDF("normal," z).

Index

D

E

F

M

Q

R

S

T

U

V

W

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.